Universitetet i Oslo
Institutt for Lingvistiske og Nordiske studier

Automatisk gjenkjenning av norske kollokasjoner

Eszter Horvati

Masteroppgave i IT-Språk, Logikk, Psykologi (Humanistisk Informatikk)

29. mars 2005

Forord

Denne oppgaven er det siste leddet i fullføringen av en mastergrad i IT-Språk, Logikk, Psykologi ved Institutt for Lingvistiske og Nordiske studier. Det sentrale temaet er statistiske prosesser i behandling av naturlige språk. Feltet ligger i skjæringspunktet mellom datalingvistikk og leksikografi.

Jeg vil takke Jan Tore Lønning som har veiledet meg gjennom planleggings- og skriveprosessen. Hans tålmodighet og oppmuntrende ord har kommet godt med under arbeidet og reddet meg fra nedturer underveis. En kjempetakk rettes også til Lars Nygaard for uvurderlig hjelp med tekniske verktøy og en alltid åpen dør. Uten Lars hadde denne oppgaven tatt mye lenger tid og krevd massevis av blod, svette og tårer.

Jeg vil også rette en aldri så liten takk til gjenværende hovedfagsstudenter og venner på instituttet som ikke har latt meg skulke unna jobben, men lokket meg med lange kaffe- og middagspauser og felles sukking... Takk Gordana, Rolf og Elise og dere på Hundremeter'n! Og ellers til alle dere som har holdt ut med klaging og syting i flere måneder nå. Jeg lover å forbedre meg!

Innhold

1 Innledning **6**

1.1 LOGON 7

1.2 Oppgavens oppbyggning 7

2 Hva er egentlig kollokasjoner? **9**

2.1 Bakgrunn 9

2.1.1 Distribusjonell tilnærmingsmåte 9

2.1.2 Intensjonal tilnærmingsmåte 10

2.2 Avgrensning 11

2.3 Klassifisering 12

2.4 Anvendelsesområder 14

3 Statistisk prosessering **17**

3.1 Korpus 18

3.1.1 Tilgjengelige ressurser 18

3.1.2 Telling av ord 19

3.1.3 Zipfs lov 20

3.1.4 Frekvens av samforekomster 20

3.2 Assosiasjonsmål 22

3.2.1 Nullhypotesen 23

3.2.2 Ensidige og tosidige tester 24

3.2.3 Eksakt hypotesetesting 24
3.2.4 Asymptotisk statistisk hypotesetesting 24
3.2.5 Punktestimering 25
3.3 NSP 25
3.3.1 Telling av n-grammer 26
3.3.2 Assosiasjonsmål for n-grammer 27
3.3.3 Filtrering 29

4 Implementasjon **30**
4.1 Lokale innstillinger 30
4.1.1 Lemma eller fullformer? 31
4.1.2 Tokenfiltrering 32
4.1.3 Typefiltre 33
4.1.4 Problemer 34
4.2 Valg av assosiasjonsmål 35
4.2.1 Kjikvadrattesten 35
4.2.2 Punktvis gjensidig informasjon 35
4.2.3 Dice-koeffisient 36
4.2.4 Phi-testen 36
4.2.5 Venstresidig Fisher-test 36
4.3 Trigrammer 37
4.4 Databasen *koll* 38

5 Evaluering **39**
5.1 Tolking av testresultatene 39
5.1.1 Resultater fra lemmatisert korpus 40
5.1.2 Resultater fra fullformskorpus 41
5.2 Er objektiv evaluering mulig? 42
5.2.1 Presisjon og funnrate 43

5.2.2 Ikke-interpolert gjennomsnittspresisjon 44
5.2.3 Manuell annotering 45
5.3 Sammenligning av kontekststørrelser 46
5.4 Forslag til forbedringer 47

6 Konklusjon **48**

A Utdrag fra korpus **50**

B Utdrag fra rangeringer **52**
B.1 Direkte søk 52
B.1.1 Fullformer 53
B.1.2 Lemmaer 59
B.2 Absolutte rangeringer 65

Kapittel 1

Innledning

Det får da være grenser!

På skolen lærte vi at språket består av forskjellige typer "byggeklosser"; bokstaver som danner ord, ord som danner setninger og setninger som danner hele bøker. Det hørtes enkelt ut. Helt til vi måtte lære oss nye språk. Da oppdaget vi plutselig at oppgaven var mer komplisert enn det først så ut til. Det var jo ikke bare å slå opp hvert eneste ord i ordboka og skrive ned betydningen! Hvorfor ikke det? Fordi det viste seg at språk består av flere type "byggeklosser" enn bokstaver og ord. Det finnes også noen enhetlige størrelser som går på tvers av ordenes grenser. De dannes av to eller flere ord, og kan ikke brytes opp. Hvor går egentlig grensene for slike enheter? Hva kan vi gjøre for å fastslå disse grensene med automatiske metoder?

Datalingvistikk er feltet som benytter seg av datateknologi for å manipulere elementer av naturlige språk. Forskere innen dette feltet ønsker å avdekke sannheter om språkets struktur og avgrensning for å komme nærmere en forståelse av det. Det er noe ved språk som fascinerer oss til stadighet. Det meste av utfordringen for en datalingvist består i å få en maskin til å *forstå* språklige enheter på samme måte som vi selv forstår dem. Problemet er bare at maskiner er som "sosialt uintelligente" mennesker: de kan ikke lese mellom linjene. En datamaskin må fortelles *nøyaktig* hva den skal gjøre med et spesielt ord eller en frase, hvordan den skal behandle dem, hvilken betydning de skal få. Den må få algoritmiske instruksjoner for hver eneste lille operasjon den skal foreta, noe som forutsetter også at vi *vet* hvilke lover og regler språket fungerer etter. Vi må kunne støpe språkstrukturen i algoritmer. Det er her vi møter på hindringene. Språkets plastisitet og dynamiske natur gjør at mange fenomener ikke kan forklares med uttalte grammatiske regler. Hvorfor noen enheter er tettere knyttet sammen enn andre, hvorfor visse ord bare kan forekomme i spesielle konstruksjoner er særegenheter ved språket

som er så å si umulig å klassifisere. Denne oppgaven er et forsøk på å samle informasjon om slike fenomener ved hjelp av telling av ord, det vil si, statistiske metoder. Det overordnede målet er å identifisere lingvistiske enheter i norsk som består av to eller flere ord, såkalte kollokasjoner. Blant dem er vi spesielt interessert i kollokasjoner som ikke lar seg oversette ord for ord til et annet språk. Disse kalles ikke-komposisjonelle enheter.

I "store" språk som engelsk og tysk har allerede mye forskning vært gjort for identifiseringen av slike enheter som går på tvers av ordgrensene. Mange datalingvister har lenge vært opptatt av fenomenet som på engelsk har fått navnet "collocations". Det har så langt ikke vært stor aktivitet innen dette feltet i Norge. Mye av motivasjonen for å identifisere kollokasjoner stammer fra leksikografisk arbeid. Konstruksjon av en kollokasjonsordbok til bruk i norsk språkundervisning for fremmedspråklige er et eksempel på konkret anvendelse av utvunnet informasjon. Samtidig ønsker man å utvikle ressurser som identifiserer kollokasjoner for å integrere dem i større datalingvistiske applikasjoner, for eksempel til bruk i maskinoversettelse.

1.1 LOGON

Denne oppgaven er tilknyttet det omfattende forskningsprosjektet LOGON ved Institutt for Lingvistiske og Nordiske studier. Prosjektets mål er å utvikle en maskinell oversetter som tar inn norsk tekst og returnerer oversatt engelsk tekst. Systemet er delt i flere moduler innen analyse, "transfer" og generering. Det ble tatt sikte på å utvikle en demonstrator for oversettelse fra norsk til engelsk i første omgang for å få størst mulig nytteverdi. Kun bokmål regnes som kildespråk, nynorsk ble ikke trukket inn for å komme lengst mulig i dybden. I den leksikalske behandlingen i analysemodulen er det ønskelig om man har integrert et verktøy som undersøker om det finnes enheter bestående av flere ord som ikke kan oversettes direkte til målspråket. Leksikalske ressurser kan så "flettes" til å gi en mest mulig helhetlig tilgang til analyse av norske tekster. Min oppgave tar sikte på å legge grunnlaget for utviklingen av en database for kollokasjoner ved hjelp av et statistisk identifiseringsverktøy.

1.2 Oppgavens oppbyggning

I kapittel 2 gjennomgås grunnbegrepene som forskning rundt kollokasjoner bygger på. Vi ser på forskjellige forståelser og avgrensninger av fenomenet kollokasjon. Definisjonene er i stor grad basert på Stefan Everts (2004) doktorgradsavhandling og John Sinclairs (1999) artikkel "The Computer,

the Corpus and the Theory of Language".

Kapittel 3 argumenterer for at statistisk prosessering er en gangbar vei mot forståelsen av fenomener i naturlige språk. Vi ser på tilgjengelige ressurser spesielt for norsk, og beskrivelser av grunnleggende statisktiske begreper. De forskjellige metodene for måling av assosiasjon mellom ord i umiddelbar nærhet til hverandre blir diskutert. Vi presenterer også systemet som brukes i oppgaven, Ngram Statistical Package.

I kapittel 4 drøftes de forskjellige valgene som må tas i anvendelsen av det statistiske verktøyet. Hvilke beslutninger man lander på i forhold til behandlingen av de tilgjengelige tekstressursene er avgjørende for beregningenes resultater. Vi ser også på problemer som oppstår som resultat av unøyaktigheter i forprosesseringen.

Kapittel 5 presenterer noen av resultatene til systemet. Vi ser på hvilke former for evaluering som kan benyttes, og i hvor stor grad disse er objektive. Beregning av presisjon og ikke-interpolert gjennomsnittspresisjon blir brukt i disse evalueringene. Muligheter til fremtidige forbedringer blir også diskutert. Til sist i oppgaven er fokus på hvilke konklusjoner som kan trekkes fra arbeidet. Databasens anvendelsesområder og integrasjonsmuligheter drøftes kort.

Kapittel 2

Hva er egentlig kollokasjoner?

2.1 Bakgrunn

Det er for tiden blitt vist stor interesse blant lingvister og leksikografer for fenomenet som på fagspråk kalles kollokasjon. Begrepet har vi fra den britiske forskeren John Rupert Firth som var en av de største aktørene for å gjøre lingvistikk til en selvstendig forskningsdisiplin i Storbritannia på 1940-tallet. Hans definisjon av mening vakte stor oppsikt hos Firths samtidige. Han mente at mening var en funksjon eller en effekt av en entitet i en spesiell kontekst. På det leksikalske plan innebærer det at ord får mening ved å stå i kontrast og relasjon til andre enheter innenfor den samme omgivelsen. Denne teorien gir opphav til begrepet om kollokasjoner, eller "vanemessige ordkombinasjoner" som Firth kalte dem. En kollokasjon er, ifølge dette synet, et uttrykk der to eller flere ord danner en enhet som er den konvensjonelle måten å uttrykke noe på. Firths velkjente utsagn "You shall know a word by the company it keeps!" (Firth 1957, s.179) har nærmest blitt et slogan for lingvister som har interesse i leksikografi og forskning på forholdet mellom samforekommende ord.

Selv om Firth skrev mye om kollokasjoner, kom han aldri frem til en klarere definisjon av hvordan begrepet egentlig kan avgrenses. Denne uklarheten har resultert i mye forvirring og uenighet i behandlingen av termen kollokasjon. Vi kan dele de motstridende synene i to hovedgrupper (Evert 2004): den distribusjonelle og den intensjonelle tilnærmingen.

2.1.1 Distribusjonell tilnærmingsmåte

Den distribusjonelle tilnærmingen er det særlig Firths etterfølgere i Storbritannia som står for. Derfor kalles den også den neofirthianske linjen. Firths

tilhengere betraktet kollokasjoner som en direkte observerbar enhet til bruk for deskriptive formål. Sinclair (1991) gir en definisjon som er betegnende for dette synet:

> "Collocation is the occurrence of two or more words within a short space of each other in a text. [...] Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated."

Som vi vil se, er tekstprosessering ved hjelp av statistiske metoder oftest basert på denne definisjonen. Den distribusjonelle tilnærmingsmåten omfatter frekvensinformasjon og tolkninger av denne informasjonen som en indikator på graden av assosiasjon mellom ordene.

2.1.2 Intensjonal tilnærmingsmåte

Den intensjonale tilnærmingsmåten bruker termen kollokasjon om en mengde leksikalske fenomener. Kollokasjoner plasseres som regel i gråsonen mellom helt faste uttrykk (idiomer) og frie ordkombinasjoner. Tilhengere av dette synet vil hevde at enhver kollokasjon består av en *base* og et eller flere *kollokater*. Basen er det frie elementet, "oppslagsordet", som beholder sin opprinnelige mening også innen en kollokasjon, mens kollokatet bestemmes av basen og kan derfor bli tilordnet en ny mening i disse tilfellene. Sammen gir basen og kollokatet et intensjonalt definert konsept, uavhengig av frekvensdata fra korpus. Evert (2004) definerer kollokasjoner innenfor dette synet på følgende måte:

> "A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon."

Under Chomskys innflytelse hadde den lingvistiske forskningen, fram til de senere årene, ikke viet særlig oppmerksomhet til fenomenet kollokasjon. Chomskianerne mener at leksikonet hovedsaklig er en liste av ombyttbare ord. Dersom det forekommer restriktive ordkombinasjoner som ikke kan gjøres rede for innenfor de syntaktiske rammene, forklarer de det med såkalte "selective restrictions", dvs. preferanser på et konseptuelt plan.

2.2 Avgrensning

I datalingvistisk bruk har ordet "kollokasjon" fått en bredere og mer generell betydning. I litteraturen brukes termene kollokasjon, samforekomst ("cooccurrences" hos Evert), flerordsuttrykk ("multi-word-unit" hos Schone og Jurafsky), n-gram, idiom, leksikalske fraser eller ordpar om hverandre som betegnelse på et og samme fenomen. Uansett navn, her inngår alle kombinasjoner av ord som på en eller annen måte må sees på som enheter på tvers av ordgrensene. Enheter der meningen av *ordkombinasjonen* ikke kan sluttes fra enkeltordenes mening hver for seg. Det er den ofte gjentatte forekomsten av visse kombinasjoner av ord som gir opphav til kollokasjoner. Siden disse kombinasjonene er enheter som eksisterer selvstendig, må de også behandles deretter. Idiomer, ordtak, klisjeer, tekniske termer og stivnede uttrykk er alle forskjellige instanser av kollokasjoner. På engelsk tar man også spesielt hensyn til sammensatte ord (f.eks. *work permit*), men dette er ikke aktuelt på norsk av den enkle grunn at vi ikke skiller komponentene i sammensetningen med mellomrom (*arbeidstillatelse*). Vi burde ikke det, i alle fall.

Problemet med denne vide definisjonen av kollokasjoner er nettopp det: at den er altfor vid. Vi må ha klare avgrensninger for hva vi skal "være på jakt etter". Det er lettere sagt enn gjort. Det som nettopp gjør studiet av naturlige språk så fascinerende er at det er uendelig mange muligheter man kan kombinere ord på. Men det er også det som er grunnen til at det er vanskelig å sette klart definerte skillelinjer mellom de forskjellige fenomenene. Uttrykk som i den ene situasjonen står som en selvstendig meningsbærende enhet, vil i en annen kontekst formidle noe annet, selv om det er akkurat samme ordsekvens. Et slikt eksempel er uttrykket *å stå i*:

(1) Direktøren står i med sekretæren.
(2) Hun står i gjørme til knærne.

I den første setningen bruker vi *står i* i betydningen "har et forhold til", altså i en overført betydning. I det andre eksempelet tar man ordene i helt bokstavelig forstand. Det handler om å fysisk "være i oppreist stilling, hvilende på føttene"[1]. Hvordan skal vi kunne skille mellom disse to betydningen av samme frase?

Manning og Schütze (1999) har formulert tre punkter som kan tjene som kriterier for å avgrense kollokasjoner:

Ikke-komposisjonalitet: En kollokasjons mening er ikke komposisjonen av meningen av delene den er satt sammen av. Et eksempel på dette

[1] Kilde: Norsk illustrert Ordbok 1993

er uttrykket *ta beina fatt* som betyr "å begynne å gå". Vi hadde ikke kunnet slutte oss til denne meningen ved kun å se på betydningen til *ta, beina* og *fatt* hver for seg. Sammen danner de en ny betydning.

Uerstattelighet: Et av elementene i en kollokasjon kan ikke uten videre erstattes med et synonym eller et "likeverdig" uttrykk. På norsk sier vi *å kaste opp*. Kaste kan ikke erstattes med synonymet *hive* uten at meningen blir forandret. Eksempler på at preposisjoner ofte har helt spesielle roller i kollokasjoner er uttrykkene *å ta hensyn* til eller *av gårde*. Verken *til* eller *av* kan byttes ut med noen annen preposisjon, til tross for at noen preposisjoner kan ha tilnærmet lik betydning.

Umodifiserbarhet: I mange tilfeller er kollokasjonene så rigide at de ikke tillater noen form for modifisering av ordenes form eller plassering. Dette er spesielt tydelig i fagtermer og lånte uttrykk som *deus ex machina*, og også i idiomatiske fraser og ordtak som *i hui og hast* eller *brent barn skyr ilden*.

Der minst et av disse kriteriene oppfylles, er det stor sannsynlighet for at vi har med en kollokasjon å gjøre. Det enkleste er å fokusere på ikke-komposisjonelle uttrykk. Dersom en setning er komposisjonell, medfører det også at man finner alle ordene i deres "rette" betydning i et oppslagsverk. Man har ikke behov for å føre opp et komposisjonelt uttrykk i en ordbok som eget oppslag fordi delenes mening utgjør også den logiske meningen til helheten. Hvorvidt det er nødvendig at alle kriteriene oppfylles eller bare en eller to av dem er avhengig av den enkelte datalingvist og hennes formål med å gjenkjenne kollokasjoner. Det finnes så mange gråsonetilfeller at man blir tvunget til å gi rom for unntak fra kriteriene. Den sentrale retningslinjen i identifiseringen av en kollokasjon vil til syvende og sist koke ned til spørsmålet om uttrykket bør listes som eget oppslag i en ordbok eller ikke, hvilket igjen innebærer at det hele er et spørsmål om intuisjon. Intuisjon kan i noen tilfeller være det eneste grunnlaget for å identifisere underliggende regelmessigheter i språket. Den beste tilgjengelige metoden er å samle data fra store tekstmengder, såkalte korpus, og danne et regelverk gjennom observasjon av gjentatte hendelser.

2.3 Klassifisering

Innenfor spekteret av forskjellige kollokasjoner er det vanlig å skille mellom *frie* og *faste* kollokasjoner (Sinclair 1999), også kalt *fleksible* og *rigide* kollokasjoner (Smadja, McKeown og Hatzivassiloglou 1996). Med frie kollokasjoner menes de ordkombinasjonene der et av elementene er gitt, mens

det/de andre kan variere. Faste kollokasjoner er de "forutbestemte" fraser der ingen variasjon er tillatt.

Denne distinksjonen kan muligens tjene som en innledende klassifisering, men den gir ikke rom for alle de "semirigide" konstruksjonene som naturlige språk faktisk tillater. Frie kombinasjoner er aldri helt frie i den betydningen at dets elementer kan flyttes vilkårlig rundt og likevel beholde samme mening. De er konstruksjoner som følger språkets syntaktiske regler og semantiske utvelgelse. Uttrykkene kan kun kalles frie innenfor disse rammene. Faste kollokasjoner på den annen side opptrer under helt andre forutsetninger. Strukturen i uttrykk som *øvelse gjør mester* eller *kreti og pleti* tillater ikke at noen av elementene blir forandret på eller byttet om. Vi kan ikke si *pleti og kreti* i betydningen "hvem som helst" (selv om uttrykket visstnok kommer fra hebraisk og betyr "kretere og filistere", noe som for så vidt kunne vært en ombyttelig koordinasjon). Det bør også nevnes at vi har såkalte stivnede uttrykk som *til sjøs* eller *til bords* som er rester etter en kasusbruk vi ikke lenger har på norsk. I en del dialektuttrykk ser vi eksempel på at gamle kasusformer har blitt bevart i stivnede uttrykk, som i *mæ gutom*.

Vi har altså (noen få) faste kombinasjoner som bare kan listes opp og identifiseres, mens de frie kombinasjonene kan gi opphav til en nærmest ubegrenset mengde variable uttrykk. Utfra dette er det tydelig at distinksjonen mellom frie og faste kollokasjoner ikke er tilfredsstillende for å gjøre rede for hele segmentet av spesielle ordkombinasjoner vi er interessert i. Sinclair (1999) foreslår å innføre en kontinuerlig skala av variable uttrykk istedenfor å etablere en kontrast mellom frie og faste kollokasjoner. Han mener at alle kollokasjoner kan plasseres et sted i dette kontinuumet der de omtalte tilstandene utgjør ytterlighetene. Forskjellen mellom variable uttrykk og frie kollokasjoner er, ifølge Sinclair, vanskelig å beskrive på grunn av et mangelfullt teknisk begrepsapparat. Likevel hevder han at de frie kollokasjonenes syntaktiske konvensjoner kun bestemmes av grammatisk struktur, mens variable uttrykk styres av ordenes kombinatoriske egenskaper.

Sinclair (1999, s.26) presenterer et forslag til klassifisering av de forskjellige variantene av ofte forekommende ordkombinasjoner som han kaller *leksikalske objekter* ("lexical item"). Denne inndelingen bygger på forståelsen av et leksikalsk objekt som den overordnede enheten i et leksikalsk hierarki. Et leksikalsk objekt består av ett eller flere ord som settes sammen etter visse syntagmatiske konvensjoner, både leksikalske og grammatiske. Disse enhetene klassifiseres i fem grupper fra de mest rigide mot de frie:

Kjernen formes av de uforanderlige elementene, denne realiseres som regel i et idiom som:

(3) Han kom til møtet *i grevens tid*.

Kollokasjoner er resultatet av en utvelgelse av ord som ikke nødvendigvis forekommer sammen med kjernen, men har en tendens til å gruppere seg rundt den ("clustering"). I eksempelet under fungerer *oppmerksomhet* som kjerne som heller "tiltrekker" seg ordet *vie* enn f.eks. *gi*:

(4) Prosjektet vil også *vie oppmerksomhet* til det vitenskapelige personalets tid.

Kolligasjoner er fenomenet der kjerneordet ikke er det viktigste, men dets grammatiske klasse. Typisk form for denne typen leksikalske fraser er *verb + refleksivt pronomen*.

(5) Hun *tok seg i å* ønske et lite øyeblikks tap av besinnelsem.

Semantisk preferanse bygger kun på semantiske kriterier, både kjerneordet og ordklassen kan variere, det gir en indikasjon om fremhevingen av samtaleemnet

(6) Forsvaret befinner seg *på konkursens rand*.

Semantisk prosodi tjener talerens kommunikative intensjon med en pragmatisk orientering, betoningen og intonasjonen kan ha negativ eller positiv retning – eksempel på negativ prosodi er ordet *begå* som ofte assosieres med ordene *selvmord* eller *forbrytelse*:

(7) Det er straffbart å *begå kriminelle handlinger*.

Grunnantagelsen til Sinclair er at det er de leksikalske objektene som er de viktigste byggestenene i språket, ikke strenger av ord. I definisjonen av de leksikalske objektene inngår både semantikken og de grammatiske relasjonene uten restriksjoner fra ytre grammatiske enhetsgrenser. Først etter at de leksikalske objektene er identifisert, skal man se etter grammatiske relasjoner. Dette kan sies å være en *fra-skallet-og-inn* tilnærmingsmåte til beskrivelsen av naturlige språk. Jeg skal ikke her ta stilling til hvorvidt dette er en god teori.

Sinclairs klassifisering av de leksikalske objektene danner et godt utgangspunkt for evalueringen og systematisering av de empiriske funnene. Jeg vil for enkelthets skyld bruke betegnelsen kollokasjon gjennomgående i oppgaven i samme betydning som Sinclair bruker samletermen leksikalsk objekt.

2.4 Anvendelsesområder

Hva skal vi gjøre med kollokasjonene dersom vi engang har klart å finne dem? Å identifisere kollokasjoner kan være nyttig på mange måter. Hvis man

etterstreber en bedre forståelse av naturlige språk og deres særegenheter, er det naturlig å måtte undersøke eksistensen av leksikalske objekter også. Distribusjonelle kollokasjoner representerer de *observerbare* bevisene som kan hentes ut direkte fra et korpus. Med enkle automatiske prosesser skiller vi ut ord som ofte forekommer sammen i naturlig tekst. Disse dataene må deretter bearbeides, dvs. generaliseres over ved hjelp av statistiske metoder for deretter å kunne anvende dem til å forutsi hvilke ordkombinasjoner som med størst sannsynlighet vil opptre i andre tekster. Konkret kan kollokasjonsdata i henhold til denne prosessen anvendes:

- i oppgaver med entydiggjøring av flertydige ord og uttrykk, f.eks. i preposisjonsfraser eller flertydige analysetrær;
- for identifisering av setnings- og konstituentgrenser, f.eks. der punktum kan være både endemarkør, men også tegn på en forkortelse (Kiss og Strunk 2002);
- for å foreta visse leksikalske valg i generering av språk;
- for å forutsi menneskelige assosiasjonsmønstre fra psykologiske eksperimenter (Rapp 2002).

I tillegg danner også distribusjonelle kollokasjoner grunnlag for å sammenligne et gitt ords kollokasjonsprofil ("cooccurrence profile") med andre ords profiler ved hjelp av en vektor for assosiasjonsgradene. Avstanden mellom to slike vektorer tolkes som en indikator på graden av semantisk likhet mellom dem. På denne måten kan resultatene brukes til å identifisere semantisk likhet mellom ord, finne synonymer (Rapp 2002) eller ekvivalente termer på et annet språk til bruk i oversettelse, eller til å samle informative fraser for automatisert generering av tekstsammendrag.

De intensjonale kollokasjonene på sin side representerer ordkombinasjonenes "iboende egenskaper". Leksikografi er det feltet der slike funn gjør seg gjeldende. I tillegg til klassiske papirversjoner av ordbøker og leksika, blir flere og flere digitale språkressurser tilgjengelige. Maskinelt lesbare ordbøker finnes nå i alle kompleksitetsnivåer, fra de enkleste ordlistene til omfattende databaser av kollokasjonsinformasjon. Identifikasjon av kollokasjoner er derfor en essensiell oppgave i leksikografisk språkteknologi:

- Kollokasjoner danner viktige enheter i både enspråklige og tospråklige ordbøker spesielt på grunn av den egenskapen at de ikke lar seg oversette ord for ord. Monolingvale eller enspråklige ordbøker må kunne presentere en vid eksempelbase av kollokasjoner som gjør det enklere for en fremmedspråklig person å lære seg korrekt bruk av termene.

Språktilegnelse består på mange måter av å pugge "ulogiske" fraser og uttrykk ved hjelp av gjentatte repetisjoner. Schone og Jurafsky (2001) er for eksempel opptatt av å finne baseordet i en kollokasjon[2] som skal fungere som en egen oppslagsenhet i en maskinell ordbok.

- Kunnskap om kollokasjoner er viktig ved automatisk generering av naturlige språk for at resultatet skal ha en naturlig flyt og ikke resultere i "unorske" uttrykk som f.eks. *gjøre en avgjørelse* istedenfor *ta en avgjørelse*.

- Syntaktisk og semantisk uvanlige ordkombinasjoner er sentrale elementer i dypsyntaktisk analyse, særlig i leksikaliserte grammatikker som HPSG og LFG.

- For maskinoversettelse er maskinelt lesbare kollokasjonssamlinger med ekvivalente oversettelser uunnværlige. Det kan være systemer basert på parallellkorpus som produserer p-ord oversettelser av n-ord kollokasjoner der p og n ikke nødvendigvis er like (Smadja et al. 1996).

Målet i denne oppgaven er å identifisere kollokasjoner fra et norskspråklig korpus med særlig vekt på ikke-komposisjonalitet som et kriterium. Det er ønskelig å kunne behandle de gjenkjente frasene som selvstendige enheter for deretter å kunne oversette dem til målspråket engelsk. Data vi utvinner ved metodene som blir beskrevet i denne oppgaven gir en rangert liste av distribusjonelle kollokasjoner som vi så kan tolke som *kandidater* til intensjonale kollokasjoner.

[2] *Baseord* er min tolkning av det forfatterne kaller "multi-word-unit headwords".

Kapittel 3

Statistisk prosessering av naturlig språk

En av de basale målene i lingvistisk forskning er å oppnå forståelse av de lingvistiske strukturene språket kommuniserer med. Man har hevdet at det finnes *regler* for å strukturere språklige uttrykk. Gjennom mange forskjellige tradisjoner har lingvister prøvd å beskrive hva som er riktig eller galt formulerte utsagn i språket. Det er bare et problem med denne forestillingen, i Edward Sapirs (1921) formulering: "All grammars leek" (gjengitt i Manning og Schütze 2003, s.3). Det er ikke mulig å gi en presis og utfyllende karakterisering av hva det er som skiller velformede setninger fra alle andre setninger som vi registrerer som grammatisk uriktige. Språkbrukere vil alltid tøye regelverket til akkurat sine behov – det er derfor språk i bruk sies å være naturlig og dynamisk. Det er selvsagt heller ikke slik at vi ikke kan gjenkjenne noen form for regelbasert kunnskap om naturlige språk. Syntaktiske regler er i høyste grad gjeldende, men de er ikke *tilstrekkelige* for å gjøre rede for språkets dynamiske egenskaper.

En alternativ tilnærming består i å angripe saken fra motsatt ende: Istedenfor å dele opp setninger i grammatiske enheter, vil vi heller spørre oss hvilke mønstre vi kan oppdage ved å undersøke den faktiske *bruken* av språk hos mennesker. Hovedredskapet i en slik undersøkelse er å *telle* ting, m.a.o. bruk av statistikk og sannsynlighetsberegning. Denne empiriske tilnærmingsmåten innebærer at man oppnår forståelse om språkenes kompleksistet og struktur ved å spesifisere en generell språkmodell, og ved bruk av statistiske mønstergjennkjenningsmetoder ("pattern recognition") kan slutte de nødvendige parameterverdiene. Det er dette synet som danner utgangspunktet i metodene jeg vil presentere i dette kapittelet.

3.1 Korpus

Den grunnleggende forutsetningen for å kunne gjøre statistiske beregninger over noe er å ha tilgang på en stor populasjon av de aktuelle forskningsobjektene. For språkteknologifeltet er dette tekst, mengder av tekst. Vi må ha så mye tekst som mulig tilgjengelig til prosessering. Praktiske hindringer gjør at forskere innen statistisk språkteknologi ikke kan bruke data om språkbruk i direkte kontekst fra virkeligheten. Isteden tyr man ganske enkelt til nedskreven tekst i store mengder, der man regner den *tekstlige* omgivelsen for kontekst.[1] Slike store samlinger av nedskreven tekst kaller vi *korpus*. Korporaene består som regel av tekster fra mange forskjellige sjangre – skjønnlitteratur, journalistikk, sakprosa, juridiske tekster, etc. – for å kunne brukes som standardreferanse om språkbruk. Takket være den raske og eksponensielt voksende teknologiske utviklingen er ikke lagring og prosessering av store mengder tekst lenger et problem. Et av de første digitale engelskspråklige korpusene er The Brown Corpus på 1 million ord, andre er International Corpus of English – Great Britain (ICE-GB) (http://www.ucl.ac.uk/english-usage/ice-gb/), The Bank of English (http://titania.cobuild.collins.co.uk/boe_info.html) og British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/). De to siste er på henholdsvis 200 og 100 millioner ord. WebCorp (http://www.webcorp.org.uk/) er et verktøy som er utviklet for å kunne bruke hele verdensvevens tekster som ett enormt korpus.[2] Dette er spesielt interessant med tanke på evaluerings- og observasjonsmetoder. Mange korpus er også bygget slik at hvert ord er annotert med grammatisk informasjon, såkalte tagger.

3.1.1 Tilgjengelige ressurser

Tekstlaboratoriet ved Universitetet i Oslo har utviklet Oslo-korpuset (http://www.tekstlab.uio.no/norsk/bokmaal/) som er en tagget samling av norske tekster på drøye 18,3 millioner ord. Det finnes i tillegg såkalte *parallellkorpora* som lister en mengde tekst med deres tilsvarende oversettelser. English-Norwegian Parallel Corpus (ENPC) og Oslo Multilingual Corpus (OMC) er slike. For oppgavens vedkommende er det tilstrekkelig med et enspråklig korpus som de statistisk funderte applikasjonene kan brukes på. Derfor har jeg valgt Oslo-korpuset som treningskorpus.

I tillegg er det nyttig å ha elektroniske ordbøker tilgjengelig. For engelsk finnes blant annet det nyttige hjelpemiddelet WordNet (http://wordnet.princeton.edu/) som er et fritt distribuert digitalt oppslagsverk til søk på både ord og fraser,

[1] Det foregår også arbeider med å samle inn data fra reell tale, såkalte talespråkskorpus. I Oslo er et slikt korpus, NoTa, under utvikling.

[2] Dog kun egnet som søkekorpus, ikke for annen type prosessering.

samt synonymer, antonymer, hyponymer etc. Tilsvarende for norsk er antakelig tjenesten Ordnett (http://www.ordnett.no, Kunnskapsforlaget AS), som samler en mengde enspråklige og tospråklige ordbøker, men denne er dessverre ikke gratis. Med tilgang til databasen til Bokmålsordboka har jeg likevel kunnet utføre søk for innhenting av informasjon for oppgavens formål.

3.1.2 Telling av ord

Den probabilistiske tilnærmingsmåten til tekstprosessering forutsetter noe forarbeid. For det første må man vite hvor ord- og setningsgrenser går. De fleste indoeuropeiske språk, deriblant også norsk, bruker mellomrom som distinktivt skilletegn mellom ord.[3] På norsk kan man også betrakte sammensatte ord som egne oppslagsenheter, ettersom de skrives i ett, uten mellomrom. Hvor mange ord finnes så i et korpus? Dette spørsmålet kan besvares på to måter. De distinkte leksikalske enhetene kalles *typer*, mens alle instanser av dem i en tekst blir referert til som *tokens*. Et *tokenisert* korpus er tekst som har markerte skiller mellom de leksikalske enhetene, eller tokens. Når man er interessert i frekvensdata for ord, teller man antall forekomster av ordformene og samler dem (vanligvis) i en rangert liste der høyest frekvens kommer først. De oftest forekomne ordene er, ikke overraskende, ord uten semantisk tyngde, såkalte funksjonsord:

```
722724  i
701854  og
522926  det
405719  er
393905  som
358789  på
358606  til
355472  en
343323  av
316634  for
311346  å
271409  at
262964  med
225423  de
222919  ikke
```

[3] I motsetning til agglutinerende språk som f.eks. ungarsk eller tyrkisk, der flere ord ofte er "klistret sammen" til å utgjøre én ortografisk enhet. I slike tilfeller må man ha en morfologisk analysator som avgjør hvor ordskillene er.

3.1.3 Zipfs lov

I boka *Human Behavior and the Principle of Least Effort* (1949) presenterer George Kingsley Zipf sin teori om at mennesker av natur alltid etterstreber minste motstandsvei (gjengitt i Manning og Schütze 2003, ss.23–27). Ifølge Zipf prøver man å minimalisere den sannsynlige gjennomsnittsraten for anstrengelsene. Dette gjelder også under produksjonen av språk, hevder han. Hvis vi teller hver ordtype i et stort korpus og rangerer dem etter frekvens på forekomstene, kan vi undersøke forholdet mellom et ords frekvens f og ordets rangering i lista r. Zipfs lov, eller det vi kan kalle en karakterisering av visse empiriske fakta, sier at:

$$f \propto \frac{1}{r}$$

Eller formulert på en annen måte: det finnes en konstant k slik at $f \times r = k$.

Zipfs lov er en omtrentlig beskrivelse av distribusjonen av ord i menneskelig språkbruk. Det viser at det finnes noen få veldig vanlige ord, et varierende antall av "mellomfrekvente" ord og mange lavfrekvente. Ifølge Zipf er dette en viktig observasjon som tyder på at både taler og mottaker prøver å minimalisere "anstrengelsene" sine; taleren ved å ha et så lite som mulig vokabular av høyfrekvente ord, mottakeren ved å lagre et relativt stort vokabular av sjeldnere ord for enklere å tyde beskjeden. Observasjonen impliserer at i prosessering av naturlig språk står vi overfor problemet med at veldig mange ord forekommer spredt i korpus. Vi vil få mange eksempler som berører kun noen få typer av ord.

3.1.4 Frekvens av samforekomster

Når vi skal finne distribusjonelle kollokasjoner, er vi interessert i mer enn enkeltordenes forekomstfrekvens. Vi vil vite noe om hvor ofte ordene forekommer sammen *i tillegg* til hvor ofte de forekommer hver for seg. Det gir liten nytteverdi i letingen etter kollokasjoner å bare se på "rå" frekvenser for to og to ord. Nedenstående liste av ordpar er de 15 mest frekvente[4], men ingen av dem kan sies å være kollokasjoner:

```
76541 det er
60537 til å
50434 for å
39288 er det
37108 at det
```

[4] Jeg har fjernet bigrammet av to punktum, som egentlig er det 6. mest frekvente paret i lista, fordi disse vil filtreres bort under videre behandling.

```
36782 det var
24355 som er
23214 i en
17938 at han
17587 at de
17383 er en
17365 og det
16347 i det
16116 var det
15806 i den
15689 som en
```

Hvis u, v er variabler for typer, vil hver type av ordpar være $x = (u, v)$. Det er altså ikke bare frekvensen $f(x)$ som er av interesse, men også de tilfellene der u og v forekommer sammen med *andre* ord. Denne informasjonen kan vi representere ved å samle frekvensverdiene av alle mulige kombinasjoner i en *kontingenstabell*. Tabellen består av 2 rader og 2 kolonner med to rader for tilfellet der u er til stede og der det ikke er til stede, og tilsvarende to rader for v tilstedeværelse eller fravær. Klassifikasjonen gir fire observerte tallverdier: O_{11}, O_{12}, O_{21} og O_{22}, som i tabellen under. (Stor bokstav U, V brukes for å betegne en stokastisk variabel, mens de små bokstavene u,v betegner en bestemt verdi i verdimengden.)

	$V = v$	$V \neq v$
$U = u$	O_{11}	O_{12}
$U \neq u$	O_{21}	O_{22}

Tabell 3.1: Kontingenstabell av observerte frekvenser

Et konkret eksempel i tabell 3.2 er forekomstdataene til ordparet $x = (til, nd)$. Disse talldataene er hentet fra Oslo-korpuset. Det er 40 instanser av x, 193 tilfeller der *nød* er andre ord i et ordpar med et annet ord enn *til* som det første ordet, og 193 847 tilfeller der *til* er første ord og *nød* ikke er andre. Det er tilsammen 11 477 737 par som hverken inneholder *til* eller *nød*.

Når frekvenser av samforekomster fremstilles i kontingenstabeller, er det vanlig å inkludere også rad- og kollonnesummeringene. Disse representeres av henholdsvis R_1, R_2 og K_1, K_2 i tabell 3.3 under. Rad- og kollonnesummene kalles *marginale frekvenser* fordi de står i tabellens marg. C står for hele utvalget, i dette tilfellet korpusets størrelse.

	$V = nd$	$V \neq nd$
$U = til$	40	193 847
$U \neq til$	193	11 477 737

Tabell 3.2: Kontingenstabell for ordparet "til nød" i Oslo-korpuset

	$V = sted$		$V \neq sted$	
$U = av$	O_{11}	+	O_{12}	$= R_1$
	+		+	+
$U \neq av$	O_{21}	+	O_{22}	$= R_2$
	$= K_1$	+	$= K_2$	$= C$

Tabell 3.3: Kontingenstabell med rad- og kolonnesummer

3.2 Assosiasjonsmål

"Rå" frekvensdata gir oss noe informasjon om ofte gjentakende ordkombinasjoner, men ikke nok til å kunne si oss noe om hvilke "bånd" som eksisterer mellom ordene partypene er satt sammen av. Disse tallverdiene er vanskelige å tolke, og de vil i alle fall ikke kunne gi noe kunnskap om annet enn det aktuelle datasettet de er hentet ut fra. Dersom hvert av ordene som utgjør et ordpar hver for seg forekommer like frekvent i korpuset, er det ingenting som tilsier at deres samforekomst ikke kun skyldes tilfeldigheter. Det er nødvendig å opprette en modell som *tolker* informasjonen vi får fra å telle, vi ønsker å måle graden av "tiltrekningskraft" eller assosisasjon mellom komponenetene i ethvert ordpar. Man kan utfra dette hevde at statistisk analyse har tre hovedoppgaver: (i) å tolke observerte frekvensdata som en indikator på *statistisk assosiasjon* mellom ord, og rangere denne graden av assosiasjon; (ii) å kunne gi en generell modell *utover* det begrensede tekstkorpuset det hentet data fra; og (iii) å filtrere ut "støy" som automatiske forprosesseringsmekanismer (tagging, parsing etc.) har introdusert. Den underliggende logikken bak alle tre oppgavene er at graden av statistisk assosiasjon mellom komponentene i en partype er latent i kontingenstabellen av observerte frekvenser. Formelen som beregner assosiasjonsverdien ("association score") utfra frekvensinformasjonen i en kontingenstabell kalles assosiasjonsmål. Assosiasjonsverdien angir hvor sterk assosiasjonen er mellom ordene, slik at

ordparene i datasettet deretter kan rangeres.

De fleste assosiasjonsmålenes beregninger er basert på sammenligningen mellom observerte frekvenser, O_{ij}, og forventede frekvenser, E_{ij}. Forventede frekvenser regnes ut med grunnlag i data fra tabell 3.3 på side 22. Tabellene 3.3 og 3.4 inneholder i så måte all informasjon som er nødvendig til beregningene av de forskjellige assosiasjonsmålene.

	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 K_1}{C}$	$E_{12} = \frac{R_1 K_2}{C}$
$U \neq u$	$E_{21} = \frac{R_2 K_1}{C}$	$E_{22} = \frac{R_2 K_2}{C}$

Tabell 3.4: Kontingenstabell av forventede frekvenser

3.2.1 Nullhypotesen

Vi er interessert i den statistiske assosiasjonen mellom komponentene som utgjør et ordpar. Siden denne assosiasjonen er en egenskap ved populasjonen, altså språket, er målet å trekke slutninger om det på bakgrunn av informasjon fra et observert utvalg, korpuset. Det er likevel ikke åpenbart på hvilken måte vi kan måle *graden* av assosiasjon innen en partype. Det er faktisk mye enklere å spørre seg: I hvilke tilfeller *vet* vi at det ikke finnes noen assosiasjon overhodet mellom ordene?

Prinsippet for statistisk uavhengighet sier at to hendelser A og B er uavhengige hvis informasjonen om at hendelse B har inntruffet *ikke* påvirker sannsynligheten for at hendelse A skal inntreffe. Sannsynligheten er altså den samme, uavhengig av om vi har informasjon om B eller ikke. I formelle termer:

$$P(A, B) = P(A)P(B)$$

Vanligvis ønsker vi å teste en arbeidshypotese H_1 som baserer seg på en påstand som krever bevis. Hvis vi skal teste på assosiasjon, kan det være nyttig istedet å sette opp en nullhypotese H_0 som holder hvis vi ikke har bevis for H_1. Tvilen kommer nullhypotesen til gode - den er sann inntil det motsatte er bevist (Løvås 2004). Hvis vi setter opp nullhypotesen som holder dersom to stokastiske variabler er uavhengige av hverandre, vil et variabelpar der nullhypotesen ikke slår til antas å være forbundet på en eller annen måte.

3.2.2 Ensidige og tosidige tester

Assosiasjonsmålene kan være en- eller tosidige avhengig av om de skiller mellom *postitive* og *negative* assosiasjoner, eller ikke. Assosiasjonen sies å være positiv hvis komponentene i et par forekommer sammen oftere enn hvis de var uavhengige av hverandre; negativ hvis de forekommer sjeldnere. Ensidige tester indikerer positiv assosiasjon ved høy resultatverdi, og intet belegg for assosiasjon i det hele tatt ved lav verdi (kan være både uavhengighet eller negativ assosiasjon). I motsetning gir tosidige tester høy verdi dersom det finnes belegg for sterk assosiasjon, både *positiv og negativ*. Lave verdier indikerer nesten eller total uavhengighet, uansett fortegn. Den absolutte resultatverdien i en tosidig test avhenger av assosiasjonens styrke med verdier nært 0 som indikasjon på uavhengighet.

3.2.3 Eksakt hypotesetesting

Noen av assosiasjonstestene har som mål å finne belegg for at det finnes mer enn uavhengighet i en gitt partype, dvs. at de prøver å motvise nullhypotesen. Dette tilsvarer vanlig statistisk hypotesetesting. Evert (2004) kaller slike tester for måling av assosiasjonens signifikanssansynlighet ("significance of association"). Eksakt statistisk hypotesetesting ("exact statistical hypothesis testing") er et slikt assosiasjonsmål. Utgangspunktet i hypotesetestingen er å fokusere på sannsynligheten for å gjøre en forkastningsfeil, såkalt type I-feil[5]. En feil av type I er å forkaste nullhypotesen hvis den egentlig er riktig. Eksakte hypotesetester finner signifikanssannsynligheten, dvs. sannsynligheten for graden av forkastningsfeil vi er villige til å akseptere, ved å finne frem til alle hypotetiske kontingenstabeller hvor utfallet er *høyst like sannsynlig som det observerte resultat*. Signifikanssannsynligheten er da bestemt som summen av sannsynlighetene for disse tilfellene, altså de faste marginalsummene i kontingenstabellene. Resultatverdien av slike tester leses som belegg for å *forkaste* nullhypotesen. Jo lavere denne verdien er, desto mindre sannsynlig er det at en gitt partype som tilfredsstiller nullhypotesen vil produsere en tilnærmet lik kontingenstabell ved ren tilfeldighet. Fishers eksakttest er en av de mest omtalte eksakte hypotesetestene[6].

3.2.4 Asymptotisk statistisk hypotesetesting

Såkalte asymptotiske hypotesetester er vanligvis basert på normalfordeling og unngår eksakttestenes numeriske kompleksitet. Slike tester beregner en

[5] I motsetning til feil av type II, som også kalles godtakningsfeil fordi vi feilaktig godtar nullhypotesen.

[6] Mer om dette i neste kapittel.

testobservator ("test statistic") som indikerer hvor mye den observerte kontingenstabellen avviker fra den tabellen vi ville få, forutsatt at nullhypotesen var riktig. Definisjonen av testobservatoren er avgjørende for hvordan rangeringen av de mulige kontingestabellene vil se ut. Man beregner så verdiene ved å summere over alle kontingenstabeller som er mer "ekstreme" enn den vi har observert(Evert 2004, ss. 72–73). De mest brukte asymptotiske testene er kjikvadrattesten, ("Pearson's chi-squared test"), Student t-fordelingen ("t-score") og logaritmelikhetstesten ("log-likelihood ratio").

3.2.5 Punktestimering

Et av problemene som signifikanstester lik de ovennevnte støter på er at høy assosiasjonsverdi kan komme av både høy grad av assosiasjon mellom komponentene *eller* av at det finnes store mengder bevis (f.eks. at O_{11} har høy frekvens). Punktestimater fokuserer derimot kun på *assosiasjonsstyrke*. De er såkalte sannsynlighetsmaksimeringsestimatorer ("Maximum Likelihood Estimates", MLE), den enkleste formen for direkte slutning av assosiasjonsstyrke. I punktestimering anslår vi verdien av en parameter med en enkelt verdi. Typiske assosiasjonsmål som er punktestimater er punktvis gjensidig informasjon ("Pointwise Mutual Information"), Dice-koefficienten og Odds-testen ("Odds ratio").

3.3 NSP

Det valgte verktøyet for å hente ut ordpar som har sterk assosiasjon og kan være kandidater til kollokasjoner er et fritt distribuert program kalt *Ngram Statistics Package* (NSP)[7] (Banerjee og Pedersen 2003). N-gram er bare en annen måte å betegne samforekomster av ord på. Formelt kan man si at en n-gram er en sekvens av N tokens. I tidligere versjoner av NSP[8] var det kun mulig å analysere *bigrammer*, en sekvens bestående av to ord. I juni 2001 ble BSP viderutviklet til å behandle n-grammer, altså i prinsippet sekvenser av vilkårlig mange ord definert av brukeren selv. I praksis er det likevel kun bigram- og trigramanalyse som er mulig foreløpig. Vi har installert programmets nyeste versjon, v0.51, og undersøkt hvilke endringer som var nødvendige for å bruke det over et norsk tekstmateriale.

[7] http://www.d.umn.edu/~tpederse/nsp.html

[8] Da med navnet BSP, Bigram Statistical Package.

3.3.1 Telling av n-grammer

NSP er bygd opp av to delprogrammer: `count.pl` og `statistic.pl`. Som det fremgår av navnet, består den første delen av å telle antall n-grammer i et definert korpus. Denne kjøres over "flate" tekstfiler og genererer en ny tekstfil med en liste av alle n-grammene rangert etter synkende frekvensverdi. For hvert n-gram er den minimale frekvensinformasjonen tatt med for å kunne konstruere dets kontingenstabell. Ved å se på kontingenstabellen fra avsnitt 2.5 med marginale frekvenser (her gjentatt som 3.5), kan vi beskrive formen linjene i resultatfila fra `count.pl` på formen: $u <> v <> O_{11}\ R_1\ K_1$.

	$V = v$		$V \neq v$	
$U = u$	O_{11}	+	O_{12}	$= R_1$
	+		+	+
$U \neq u$	O_{21}	+	O_{22}	$= R_2$
	$= K_1$	+	$= K_2$	$= C$

Tabell 3.5: Kopi av kontingenstabell fra avsnitt 2.5

Fra kjøringsresultater har vi følgende eksempel:

```
11750286
reise<>tiltale<>14 6343 360
```

Dette betyr at `count.pl` fant litt over 11,7 millioner bigrammer til sammen, der den unike kombinasjonen *reise tiltale* finnes 14 ganger i korpuset, *reise* forekommer 6343 ganger på første plass i et bigram, mens *tiltale* forekommer 360 ganger på andre plass i et bigram.

Systemet gir også mulighet til å forme n-grammer som ikke består kun av etterfølgende ord. Brukeren får muligheten til å definere størrelsen på et *vindu* som består av en sekvens av k etterfølgende tokens. Verdien av k er større eller lik verdien av N. Et n-gram kan så formes av ulike N tokens så lenge alle er innenfor det samme vinduet av størrelse k. Gitt et vindu på størrelsen k og et n-gram på størrelsen N, har vi altså ${}^kC_N$ (leses: k velger N) n-grammer for hvert vindu.

I sekvensen

```
de besluttet å reise tiltale mot
```

vil vi for eksempel ha behov for å finne alle mulige bigrammer for et vindu på størrelse 3, det vil si der $N = 2$ og $k = 3$. Vi får da følgende bigrammer:

```
de<>besluttet<>
de<>å<>
besluttet<>å<>
besluttet<>reise<>
å<>reise<>
å<>tiltale<>
reise<>tiltale<>
reise<>mot<>
tiltale<>mot<>
```

Når det gjelder telling av n-grammer der $N > 2$, gjelder akkurat samme prinsipp, til tross for at det er flere tall å holde styr på. For trigram-resultatet

```
9993162
reise<>tiltale<>mot<>2 4807 288 21715 8 68 9
```

er den første verdien det totale antallet trigrammer i korpus, mens tallene etter trigrammet er de forskjellige frekvensmålene for forholdene mellom hvert token som utgjør trigrammet. Sekvensen *reise tiltale mot* forekommer nøyaktig 2 ganger i korpuset, *reise* finnes 4 807 ganger på første plass i et trigram, *tiltale* er på andre plass 288 ganger, og *mot* forekommer 21 715 ganger som det tredje tokenet i et trigram i korpuset. Deretter forekommer ordene *reise* og *tiltale* sammen som første og andre token i 8 av tilfellene, *reise* og *mot* som første og tredje token i 68 av tilfellene, og *tiltale* og *mot* som andre og tredje token i 9 av tilfellene. Det generelle formatet for større n-grammer kan så sluttes av dette. Hver frekvensverdi eller frekvenskombinasjon uttrykker antallet n-grammer som har en gitt kombinasjon av et eller flere tokens på en eller flere faste plasser. Vi har totalt 2^{n-1} mulige frekvenskombinasjoner.

3.3.2 Assosiasjonsmål for n-grammer

Programmets andre del, `statistic.pl` implementerer de forskjellige assosiasjonsmålene over frekvensdata. Det er egentlig et "rammeverk" som har som oppgave å ta som "input" n-grammer med deres frekvensdata, sende disse videre til et statistisk "bibliotek" og formatere "output" fra bibliotekene til en resultatfil. Programmet kan i så måte enkelt tilpasses nye assosiasjonsmål. NSP tilbyr beregninger på følgende assosiasjonsmål for bigrammer:

- "Dice Coefficient" (`dice.pm`)

- Fishers eksakttest — venstresidig (`leftFisher.pm`)
- Fishers eksakttest — høyresidig (`rightFisher.pm`)
- "Log-likelihood ratio" (`ll.pm`)
- Gjensidig informasjon, "True Mutual Information" (`tmi.pm`)
- Punktvis gjensidig informasjon, "Pointwise MI" (`pmi.pm`)
- "Odds ratio" (`odds.pm`)
- "Phi Coefficient" (`phi.pm`)
- Student t-fordelingen, "T-score" (`tscore.pm`)
- Kjikvadratfordelingen, "Pearson's Chi Squared Test" (`x2.pm`)

For trigram er kun "Log-likelihood ratio" (`ll3.pm`) implementert.

Formen på resultatfilene `statistic.pl` genererer er slik vi ser i utdraget under. Dette er fra rangeringsresultater etter test på punktvis gjensidig informasjon:

```
11671817
tastafon<>konvent<>1 22.4765 2 2 2
meierismøret<>sylten<>1 22.4765 2 2 2
visá<>ávis<>1 22.4765 2 2 2
HOBBY<>INTERIØR<>2 21.8916 3 3 3
nam<>nam<>2 21.8916 3 3 3
juksemaker<>pipelort<>2 21.8916 3 3 3
siselerte<>metallarbeider<>2 21.8916 2 2 3
nedgravet<>frostfritt<>2 21.8916 2 3 2
sluttede<>reiseselskaper<>2 21.8916 2 3 2
barnepsykiatriske<>behandlingshjem<>2 21.8916 2 3 2
anabole<>steroider<>2 21.8916 2 2 3
```

Øverste linja er, som før, antallet n-grammer totalt i korpus. Deretter følger numeriske verdier som kan være av forskjellig betydning alt etter hvilken test som brukes. Det første tallet er i alle listene plassen i rangeringen, deretter følger den verdien assosiasjonstesten har kalkulert for n-grammet. De tre siste sifrene er akkurat de samme som var tilknyttet hvert enkelt n-gram fra frekvenstellingen.

3.3.3 Filtrering

Å filtrere kollokasjonsdata innebærer å fjerne visse "uønskede" par av ord eller ordtyper. Vi har hovedsaklig to forskjellige former for filtrering: **tokenfiltrering**, der tokens fjernes *før* man uthenter frekvensdata fra korpus; og **typefiltrering**, der typer av ordpar tas bort *etter* at frekvensdata foreligger (Evert 2004).

Tokenfiltrering har direkte påvirkning på datautvalgets størrelse (antall instanser i korpus reduseres), og derfor også på ordtypenes frekvensdata. Denne formen for filtrering kan forstås som kun et tilleggskriterium for identifisering av tokens fra korpuset. I så fall har den heller ikke videre implikasjoner på modellens adekvathet. Likevel er det et poeng å holde seg til en systematisk metode for hvilke instanser man velger å filtrere bort på denne måten. Det er f.eks. ikke nok å si at vi ønsker å fjerne bestemte/ubestemte artikler bare fordi de produserer "uinteressante resultater". Disse ordene er jo også viktige elementer i språket og også i den strukturelle oppbyggingen av hele tekstbasen.

Under typefiltrering, på den annen side, er det nettopp slike "uinteressante" n-grammer vi ønsker å reservere oss fra. I denne prosessen fjerner man n-grammer fra datasettet uten å påvirke frekvensverdiene for de gjenværende typene (siden disse er talt i utgangspunktet). Typefiltrering brukes som oftest for å forbedre de statistiske målenes resultater, og med det også å gjøre verktøyet for gjenkjenning av kollokasjoner mer presist.

NSP gir muligheten til å enkelt inkludere filtrering i programmet. Brukeren kan blant annet lage seg en "stop-liste" over visse trekk han ønsker å ta bort før prosessering av n-grammene blir gjennomført. Funksjonen `--stop` kaller på en egendefinert fil der regulære uttrykk beskriver hvilke elementer man ønsker å fjerne. Denne funksjonen fjerner bare ord som allerede er blitt definert som tokens. Hvis man ønsker å filtrere bort individuelle instanser av ikke-token sekvenser, kan funksjonen `--nontoken` brukes til dette. På denne måten elimineres disse elementene *før* tokens i korpus telles opp.

Kapittel 4

Implementasjon

Å implementere, realisere noe i datalingvistikk innebærer som regel programmering av algoritmiske operasjoner. I dette tilfellet har jeg vært så heldig å kunne bruke et ferdig system. Implementasjonen har i praksis bestått i å gjøre vurderinger og ta avgjørelser for *hvilke* fenomener vi vil ha, *hvordan* vi vil manipulere dem og *hvor* vi vil kunne lagre dem. Programmeringsoppgavene har vært minimale, og har for det meste bestått i enkle skript for å hente ut fra korpus de elementene vi vil bruke for statistisk prosessering. Jeg har fått uvurderlig hjelp fra Lars Nygaard ved Tekstlaboratoriet til denne biten. Siden NSP er skrevet i Perl, var det naturlig å bruke det samme språket til de andre oppgavene og. Selv har jeg bare tilegnet meg helt basale Perl-ferdigheter, så Lars' solide kompetanse har kommet godt med!

4.1 Lokale innstillinger

Anvendeligheten av NSP som verktøy for kollokasjonsgjenkjenning avhenger av hvilke tilpasninger man gjør i ressursene programmet skal brukes over. Korpusformatet er et av de viktigste elementene her. Oslo-korpuset er tagget med UiOs multitagger (utviklet av Tekstlaboratoriet og Dokumentasjonsprosjektet i samarbeid), og deretter med en tagger som velger blant flertydigheter (disambiguerende tagger). Korpuset er blitt gjort om til CQP-format automatisk, fra rene tekstfiler med meta-informasjon i headeren, og fra en innholdsfortegnelse med riktig tekstidentifikator. Taggene inneholder ordklasseinformasjon og andre syntaktiske og morfologiske trekk. Vi har brukt det taggede korpuset for bokmål for å hente ut noe informasjon, og deretter laget en "strippet" tokenisert versjon av den spesielt for kollokasjonsgjenfinning. Et token defineres her som en uavbrutt sekvens av bokstaver

som blir valgt ut ved hjelp av et sett av regulære uttrykk. I utgangspunktet vil vi kun ha med tokens, derfor ignoreres tagger. Setninger separeres med en åpen linje.[1] Resultatet kalles et "strippet" korpus fordi det er en flat tekstfil av ord der all overflødig informasjon har blitt fjernet.

Det viste seg at vi måtte redusere størrelsen på korpuset til 15 millioner ord for at telling av bigrammer og trigrammer skulle gå gjennom. Dette er fortsatt en rimelig størrelse, spesielt med tanke på at vi i utgangspunktet hadde regnet med at trigramtellingene ville ta så mye plass at bare halve størrelsen av det opprinnelige korpuset ville kunne brukes. 15 millioner kan sies å være et godt kompromiss.

4.1.1 Lemma eller fullformer?

Mange foretrekker å jobbe med statistiske data fra et korpus som er lemmatisert, dvs. at alle ordene er redusert til grunnformer. Evert (2004) argumenterer for denne bruken fordi han mener at et lemmatisert korpus resulterer i betydelig forskjell i forekomstfrekvensene av n-grammer i forhold til et korpus med fullformer. Selv om dette er et godt poeng, ser jeg nytten av å undersøke visse fenomener i et fullformskorpus også. Det finnes tross alt noen uttrykk som bare brukes i spesielle, av og til arkaiske, former. Dersom ordene som utgjør slike uttrykk blir brutt ned til lemmaer, kan det bli vanskeligere å identifisere disse kollokasjonene.

Et eksempel er uttrykket *å gå god for noe/noen*. Dette kan kalles et *semirigid* uttrykk fordi deler av det er foranderlige mens andre deler må beholde sin opprinnelige form; verbet *gå* kan bøyes i tid, men adjektivet *god* kan ikke bøyes hvis uttrykkets opprinelige mening skal bli bevart:

(8) Granskingsutvalget *går god for* rapporten.

(9) *Granskingsutvalget *går bedre for* rapporten.

Ved å søke i korpuset finner vi at det er 23 tilfeller av uttrykket *gå god for* i materialet med lemmaer, mens vi kun får 16 tilslag med fullformer for *gå* og lemmaet for *god*. Avviket på 7 forekomster skyldes nettopp at lemmasøket også inkluderer andre former av adjektivet der disse konsekvent blir redusert til grunnformen *god*. Setningen

(10) Ting *går bedre for* ham nå.

er grammatisk helt korrekt, men ingen vil vel hevde at de kursiverte ordene danner en kollokasjon. Hvis vi bruker kun lemmatisert korpus, vil

[1] Dette for å unngå tellingen av n-grammer som går på tvers av setningsgrensene.

vi altså for dette eksempelet få tilslag av 7 "falske positiver", dvs. n-grammer som feilaktig identifiseres som kollokasjonskandidater.[2] Når antall n-gramforekomster telles, vil slike tilfeller resultere i en skjev fordeling av frekvensverdiene, og følgelig også påvirke statistikken av dem.

Legg forøvrig merke til at bruken av *samsvarsbøyning* av adjektivet til tider er tillatt. Språket er antagelig i ferd med å akseptere samsvarsbøyning av adjektivet i flertall i dette uttrykket. Et søk i WebCorp på frasen *går gode for* ga 28 tilslag, selv om den vanlige bruken fortsatt er *gå god for*, også i flertall. Samtidig er ikke intetkjønnsformen akseptert. Eksempler fra korpuset viser at alle forekomster av *gå godt for* er i en rent komposisjonell setting, og vi er jo i utgangspunktet på jakt etter ikke-komposisjonelle samforekomster.

På den annen side er det ikke til å komme unna at identifiseringen av de fleste n-grammer, de fleksible, blir betydelig enklere i et lemmatisert korpus. Variasjoner i ordformer reduseres i et lemmatisert materiale, og vi kan mer effektivt generalisere over flesteparten av ordkombinasjonene. Ved hjelp av informasjon fra taggene i Oslo-korpuset har vi kunnet hente ut alle ordenes grunnformer.[3] Under tokeniseringsprosessen har vi derfor generert to flattekstkorpus: ett lemmatisert og et med fullformer. NSPs delprogrammer blir så kjørt over begge tekstfilene. Forskjellene diskuteres i avsnitt 5.1.

4.1.2 Tokenfiltrering

Som første steg i tokeniseringsprosessen ble det foretatt en del filtreringsvalg. Det er dette som kalles tokenfiltrering. Ved hjelp av informasjon fra taggene i det originale korpuset har vi kunnet skille ut og markere elementer som ikke er interessante i forhold til kollokasjonsidentifisering. Tall, store bokstaver, periodeskilletegn som punktum, komma, ellipser og andre spesielle skrifttegn ble tatt bort ved enkle regulære uttrykk. I tillegg har attributter i taggene vært nyttige kjennetegn for å kunne markere egennavn (`prop`), forkortelser (`fork`), symboler (`symb`), dato og klokkeslett (`<dato>, <klokke>`), romertall (`<romertall>`). Disse elementene har blitt utstyrt med markøren `_STOPP`. Under kjøring av `count.pl` for alle bigrammer med vindusstørrelse fra 2 til 5 og trigrammer, er det spesifisert at programmet ikke skal telle tokens som er markert med `_STOPP`-tegnet. Ved hjelp av NSPs stoppfunksjon blir disse ganske enkelelt oversett.

[2] Dette forklares i detalj i avsnitt 5.2.1.

[3] Noen feiltolkninger i omforming til lemmaer forekommer på grunn av feil ved den automatiske taggeren. Flertydigheter på ordnivå som taggeren ikke har løst har vi heller ikke kunnet ta høyde for. Vi regner med at feilene gjelder en såpass liten del av materialet at det ikke påvirker opptellingene i betydelig grad.

4.1.3 Typefiltre

Vi har tidligere definert typefiltrering som den prosessen der man fjerner linjer fra allerede talte forekomster av tokens. I NSP er det en egen funksjon som fjerner forekomster under en viss grense som brukeren selv kan definere. Denne funksjonen, `--frequency`, trer i kraft allerede mens telling av forekomstene foregår. Vi har spesifisert at programmet skal fjerne alle n-grammer som forekommer kun én gang. Slike n-grammer er nemlig noen av de vanligste kildene til støy i de statistiske beregningene. De fleste assosiasjonsmålene vil betrakte n-grammer med én forekomst der også de individuelle komponentene forekommer få ganger, som kollokasjoner med høy assosiasjonsverdi. I virkeligheten er de aller færreste (om ingen) av disse samforekomstene av interesse. Metodene gir bedre resultater når disse er fjernet i utgangspunktet.

Etter å ha kjørt assosiasjonsmålene, ønsker vi å filtrere ytterligere. På grunn av taggerens unøyaktighet "sniker" det seg inn noen feil som vi vil reservere oss mot. I korpuset forekommer en mengde egennavn som ikke gjenkjennes av taggeren, noen tall kommer med, og til tider også skilletegn som av en eller annen grunn ikke har blitt markert i tidligere utvelgelse. Metoden for å fjerne disse uønskede elementene i siste instans, er å lage et lite skript som sjekker at alle komponentene i n-grammet finnes i fullformsordlista for norsk. Dette er en liste av oppslagsordene i den leksikalske databasen som taggeren og multitaggeren baserer seg på, samt oppslagsordenes alle fulle former når de er bøyd. Lista inneholder 142 280 oppslag og tilsammen 1 221 212 former av dem. Det må bemerkes at fullformsordlista inneholder allerede en mengde faste uttrykk som *til sjøs*, *av gårde* og lignende. Ettersom disse er oppslag som inneholder mellomrom, vil ingen av ordene fra n-grammene identifiseres. Siden f.eks. *sjøs* ikke forekommer som selvstendig oppslag, filtreres noen ekte kollokasjoner bort ved denne prosessen. Konkret dreier det seg om 673 uttrykk. Vi går ut fra at disse i utgangspunktet er gjenkjent og derfor ikke trenger å bli identifisert ved bruk av NSP. En liste med disse uttrykkene kunne eventuelt blitt lagt til i databasen, men det er uklart hvilke verdier de i så fall skal tilordnes.

Ulempen med denne løsningen er at vi risikerer å fjerne relevant informasjon. Det finnes en mengde sammensatte ord i fullformsordlista, men om det opptrer noen nydannelser, både enkeltord og sammensetninger, i tekstmaterialet, vil disse bli ignorert.[4] På den annen side, det vi ønsker å lage er en "normalisert" samling av potensielle kollokasjoner. Sannsynligheten for at nydannede ordsammensetninger er gode kollokasjonskandidater er rimelig liten. Ettersom utvalget i utgangspunktet baserer seg på en statisk mengde

[4]For eksempel *stillingsregulativ* er et ord som forekommer 3 ganger i korpuset, men ikke står listet i fullformsordboka.

med tekst ser jeg ikke på denne tilnærmingen som problematisk.

Vi har også vurdert muligheten for å filtrere bort "funksjonsord" som modale hjelpeverb, determinativer, eventuelt noen preposisjoner fordi disse genererer en enorm mengde n-grammer, men veldig få av betydning for kollokasjonsgjenkjennelse. Problemet med det er at det er svært vanskelig å formulere en generell regel for *hvilke type ord* vi egentlig vil ha bort. Mange preposisjoner brukes mye i funksjonelle roller, men de kan også være element i uttrykk som ikke kan oversettes ord for ord. Et slikt eksempel er *i forhold til* der ingen av preposisjonene spiller en avgjørende semantisk rolle, men de er likevel uerstattelige deler av uttrykket. På engelsk ville man sagt *compared to* som ikke er en ord-for-ord oversettelse.

4.1.4 Problemer

I tillegg til problemene som ble nevnt over, er det også andre ufullstendigheter ved utvelgelsen leseren må være oppmerksom på. Presisjonen til taggeren er en avgjørende faktor for applikasjoner som skal operere over resultatmengden som genereres ved hjelp av den. Taggeren har en leksikals funnrate ("recall") på 99 % og presisjon på 95,4 % (over testkorpus). Det er et meget godt resultat isolert sett, men der taggeren ikke gir en korrekt tolkning av lingvistisk informasjon, vil heller ikke uthentet data for kollokasjonsgjenfinning interpreteres korrekt. Dette gjelder både feil ordklasse eller ukorrekt disambiguering. Feilraten forplanter seg videre i systemet når andre applikasjoner er avhengige av resultatene til taggeren. Slike feil motiverer filtrering også i senere stadier av identifikasjonsprosessen.

Eksempler på feil som genereres som resultat av feiltagging er søket etter romertall. Allerede i preprosessering til taggingen av korpuset skal en mengde uttrykk gjenkjennes. Romertall er ønskelig å filtrere ut, men mange andre forkortelser havner i samme kategori. Under følger noen reelle utdrag fra korpuset der de uthevede elementene har fått markøren `<romertall>`:

(11) **I** 1996 inngår ni selvstendige...

(12) Av den grunn bruker **MMI** stort sett jenter...

(13) **LIV** PÅ MARS...

(14) Aktuell **CD**: "Vrimmel"...

(15) Hun ble skadet i en trafikkulykke under **VM** i Dortmund.

Korpuset er heller ikke korrekturlest. Mange skrivefeil, ortografiske feil, utenlandske ord og lignende har kommet med i teksten og blitt behandlet på lik linje med syntaktisk og morfologisk korrekt norsk tekst. Taggeren gjør en analyse av ordendelsen og tilordner ordene den(ifølge taggeren) mest

sannsynlige ordklassen. Noen ord blir tagget med markøren `ukjent`, men det er et forholdsvis lite antall, ikke en representativ mengde vi kunne basere oss på.

4.2 Valg av assosiasjonsmål

NSP muliggjør analyse av n-grammer ved hjelp av tilsammen 10 forskjellige assosiasjonsmål. Flere av disse er varianter av hverandre og kan ordnes i grupper etter hva utregningene legger vekt på. Derfor har jeg nøyd meg med å velge kun en av hver type, for at sammenligninger av målenes resultater skal være overkommelig.

4.2.1 Kjikvadrattesten

Et av de mest brukte asymptotiske hypotesetestene er Pearsons kjikvadrattest. Denne måler fordelingen mellom de observerte dataene som hadde vært å forvente dersom det første og det andre ordet var uavhengige av hverandre. Jo høyere verdi, desto mindre bevis foreligger for å konkludere med at ordene er uavhengige. Formelen under nullhypotesen for uavhengighet er som følger:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

4.2.2 Punktvis gjensidig informasjon

Gjensidig informasjon ("Pointwise Mutual Information", heretter PMI) er en metode for å sammenligne sannsynligheten for at ordene i et n-gram forekommer sammen mot sannsynligheten for at de forekommer ubetinget av hverandre (Church og Hanks 1990). Man vurderer kun ett punkt i fordelingen. Testen ser på produktsannsynligheten for et n-gram normalisert over det totale antallet n-grammer, altså:

$$pmi = log \frac{O_{ij}}{E_{ij}}$$

Dersom det finnes en reell assosiasjon mellom komponentene i n-grammet er $pmi > 0$, som også kan beskrives $P(u, v) > 0$. Hvis ingen assosiasjon er å finne er verdien tilnæremet lik 0. Ved negativ assosiasjon, dvs. når ordene forekommer sjeldnere sammen enn vi ville forventet at de gjorde hvis de var uavhengige, er $pmi < 0$.

4.2.3 Dice-koeffisient

Dice-koeffisienten er også en punktvis test, men som ikke er betinget av n-grammens forventede verdier, E_{ij}. Dette assosiasjonsmålet baserer seg kun på frekvenstallene for n-grammet og for de individuelle ordene som utgjør det. Resultatverdien er høy ganske enkelt når ordene i et n-gram forekommer oftere sammen enn hver for seg. Definisjonen formaliseres på følgende måte:

$$Dice = \frac{2 \times O_{11}}{R_1 + K_1}$$

Denne testen bruker kun de absolutte frekvensverdiene for forekomsten av n-grammets komponenter sammen og hver for seg.

4.2.4 Phi-testen

Phi-testen er egentlig en avledning fra Pearsons kjikvadrattest og beregner graden av assoisasjon mellom binære variabler. Vi får positiv assosisasjon hvis flertallet av forekomstene er å finne langs de diagonale cellene i kontingenstabellen. Fra tabellen 3.3 på side 22 ser vi at dette betyr at vi har positiv assosiasjon dersom O_{11}, O_{22} er større enn O_{12}, O_{21}, og negativ assosiasjon hvis det er omvendt. Beregningen er kun avhengig av observerte frekvenser og marginalfrekvensene:

$$\phi = \frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{R_1 R_2 K_1 K_2}}$$

4.2.5 Venstresidig Fisher-test

Fishers eksakte test er numerisk kompleks fordi beregningen går ut på å sammenligne alle kontingenstabeller som fører til identiske marginale sannsynligheter. Pedersen (1996) argumenterer sterkt for at en slik eksakttest er å foretrekke fremfor de asymptotiske testene. Testen indikerer sannsynligheten for at den observerte tabellen er hentet fra en populasjon der nullhypotesen er sann.

Fisher-testen kan tolkes som ensidig eller tosidig test. Den ensidige testen kan være høyre- eller venstresidig. I en høyresidig test summerer man de hypergeometriske sannsynlighetene for at alle tabeller med faste marginalsummer der O_{11} er større eller lik i den observerte tabellen. Den venstresidige testen er tilsvarende med den ene forskjellen at sannsynligheten for de mulige tabellene er mindre eller lik den observerte tabellens. For n-grammer viser testen hvor sannsynlig det er å se det observerte n-grammet *færre* ganger i et annet tilfeldig utvalg fra populasjon der nullhypotesen om

uavhengighet er sann. Hvis sannsynligheten er høy, viser det *avhengighet* mellom elementene i n-grammet. Summen av alle tabellenes sannsynligheter er resultatet av en tosidig test. Pedersen hevder at en venstresidig Fisher-test er mest troverdig i kollokasjonsidentifisering fordi resultatet ikke legger restriksjoner på type data den trenger. Dessuten leses den venstresidige testen på samme måte som de andre assosiasjonstestene vi bruker; jo sterkere assosiasjon, desto høyere verdi. Den høyresidige testen er symmetrisk til den venstresidige, derfor indikerer høy verdi lav eller ingen assosiasjon (fordi det er veldig sannsynlig å finne en tabell etter kriteriene i utvalget der nullhypotesen er sann).

4.3 Trigrammer

Hva skjer når vi øker n? Generelt kan man si at jo høyere verdi n har i et n-gram, desto mer presis er modellen vår. Problemet med implementasjonene av statistiske beregninger for store n-verider er at vi også får mange flere n-grammer å telle, og ikke minst å regne over. Vi har faktisk måttet redusere korpuset til halvparten fordi å regne over trigram tok veldig mye diskplass. Trigrammer representeres i en tredimensjonal kontingenstabell:

		$W = w$	$W \neq w$	Sum
$U = u$	$V = v$	O_{111}	O_{112}	R_1
$U = u$	$V \neq v$	O_{121}	O_{122}	R_2
$U \neq u$	$V = v$	O_{211}	O_{212}	R_3
$U \neq u$	$V \neq v$	O_{221}	O_{222}	R_4
Sum		K_1	K_2	C

Tabell 4.1: Trigram kontingenstabell

I NSP har vi kun ett assosiasjonsmål som beregner koblingsverdier for trigrammer, denne er realisert som logaritmelikhetstesten ("log-likelihood ratio") `ll3.pm` i programpakken. Logaritmelikhetstesten måler avviket mellom observert og forventet data forutsatt at u, v og w er uavhengige av hverandre. Høy verdi indikerer sterk assosiasjon, altså få eller ingen bevis for at ordenenes forekomst er tilfeldig. Beregningene er som følger:

$$\text{log-likelihood} = 2\sum_{ijk} O_{ijk}\ log\frac{O_{ijk}}{E_{ijk}}$$

4.4 Databasen *koll*

Vi har opprettet en database for lagring av resultatene. De forskjellige bigram-filene for samme vindu har blitt lagt i hver sin tabell, med en kolonne for hvert assosiasjonsmål. På denne måten kan man enkelt søke etter resultater rangert etter det spesielle målet man selv ønsker. Siden resultatene for bigrammer med forskjellige vindusstørrelser ikke lar seg direkte sammenligne,[5] måtte vi lage distinkte tabeller for hvert av dem. Vi har også skilt mellom fullformer og lemma. Trigrammene har på nåværende tidspunkt kun blitt telt for etterfølgende tokens, dvs. vindu 3, og har kun implementasjon for assosiasjonsmålet log-likelihood. Det vil være en fremtidig oppgave å implementere flere assosiasjonsmål og regne over videre kontekster.

Tabelloversikten er som følger:

```
+-----------------+
| Tables_in_koll  |
+-----------------+
| form_trigram    |
| form_w2_bigram  |
| form_w3_bigram  |
| form_w4_bigram  |
| form_w5_bigram  |
| lemma_trigram   |
| lemma_w2_bigram |
| lemma_w3_bigram |
| lemma_w4_bigram |
| lemma_w5_bigram |
+-----------------+
```

I neste kapittel vil vi se på hvordan resultatene i disse tabellene kan tolkes og hvilke måter man kan nyttigjøre seg dem.

[5] Se neste kapittel om evaluering av størrelse på vinduene.

Kapittel 5

Evaluering

> "How do I interpret the values of these scores?
> Carefully. Subjectively. Creatively."[1]
>
> *Ted Pedersen*

Kjernen til tolkningen og evalueringen av testresultatene ligger nettopp i disse tre holdningene; å tolke resultatene som NSP genererer er en utfordring for brukerens kreativitet og subjektive vurderingsevne. Vi skal være forsiktige med å trekke bastante konklusjoner på grunnlag av assosiasjonsmålenes resultater. Rangeringsresultatene som målene produserer er kun indikasjoner på antatte kollokasjonsegenskaper, ikke egenskaper som er direkte overførbare til virkeligheten. Vi prøver gjennom denne vurderingen også å danne oss et bilde av hvor nær relasjonen er mellom beregnet assosiasjonsverdi og reelle kollokasjonsegenskaper mellom ordene i et n-gram. I de neste avsnittene vil jeg presentere noen fakta og vurdere resultatene etter beste evne, men leseren står selv fritt til å gi andre alternative tolkninger av dem. Subjektivitet er et uunnvikelig nøkkelord i denne sammenhengen nettopp på grunn av vanskelighetene med å etablere en klar avgrensning for kollokasjoner som fenomen (se kapittel 2).

5.1 Tolking av testresultatene

Vi ønsker generelt å sammenligne assosiasjonsmålenes resultater utfra rangeringen hver av dem presenterer. Den absolutte måleverdien er ikke direkte sammenlignbar mellom testene, men alle de fem testene vi har undersøkt har det til felles at jo sterkere assosiasjon det er mellom ordene

[1] Se http://www.d.umn.edu/ tpederse/Code/FAQ.nsp-v0.67.html

i et n-gram, desto høyere verdi tilordner de denne. En av de mest intuitive måtene vi kan sammenligne testene på er ved å ta for oss ett enkelt ord og alle dets tilordninger i n-grammene. I likhet med et ordbokoppslag søker vi opp et ord for å kunne studere omgivelsene til dette ordet. Vi går ut fra *basen* og søker dets *kollokater*, slik det ble beskrevet gjennom den intensjonale tilnærmingsmåten (se avsnitt 2.1.2).

5.1.1 Resultater fra lemmatisert korpus

Jeg har valgt ordet *kaste* og sammenligner rangerte resultatlister for bigrammer med dette ordet fra de forskjellige testene. I tillegg B er det en lengre liste over direkte utdrag fra databasen av rangeringen til de forskjellige assosiasjonsmålene.

Plass	Dice	LeftFisher	Phi	PMI	X^2
1	bort	fram	smutt	smutt	smutt
2	lys	hun	loss	loss	loss
3	skygge	håndveske	blår	lp	blår
4	stein	lodd	snøball	blår	snøball
5	sig	sin	lp	snøball	lp
6	seg	sigarettstump	lys	epleskrott	bort
7	opp	inn	seg	sigarettstump	seg
8	ut	jakke	bort	PH	lys
9	loss	jeg	epleskrott	overbord	epleskrott
10	snøball	hode	sten	sneip	en

Tabell 5.1: Kollokater til *kaste* i lemmatiserte rangeringslister

I et lemmatisert korpus forekommer *kaste* sammen med *smutt* i 5 tilfeller, og kommer på første plass i hele tre av de fem rangeringene. Dette ordparet er ifølge Sinclairs klassifikasjon en ekte kollokasjon. Selv om *kaste* kan forekomme sammen med utallige andre ord, er *smutt* et ord som kun forekommer i akkurat dette uttrykket.[2] Det er ikke et idiom fordi vi kan bøye det første ordet og samtidig beholde meningen. I de samme tre testene

[2]Kaste smutt: å kaste en flat stein langs vannflate så den spretter opp igjen flere ganger (kilde: Norsk illustrert ordbok).

kommer *kaste loss* på andre plass, mens *kaste blår* følger tett på. Alle disse vil tydelig kunne plasseres i klassen for kollokasjoner. *Kaste opp* og *kaste lodd* er også kollokasjoner etter Sinclairs kriterier. *Kaste seg* vil kunne føyes under klassen for kolligasjoner, mens *snøball, epleskrott, stein* danner semantisk preferanse sammen med *kaste*.

Vi har noen tilfeller på ord som ikke akkurat kan kalles representative i dagens norskbruk. Slike er ordparet *kaste sig*, som faktisk havner høyere i rangeringen i de fleste testene enn *kaste seg*.[3] Også paret *kaste PH* vekker oppsikt. Dette bigrammet er å spore i Gerd Brantenbergs roman, "Egalias døtre".[4] Begge disse eksemplene vitner om at korpuset vi bruker ikke er ment å være representativt selv om det inneholder tekster fra ulike genrer. Hovedformålet med det er å tilby en stor tekstmengde som forskerne kan benytte til søking. Å gå ut fra at korpuset gjenspeiler dagens språkbruk ville vært illusorisk. Oslo-korpuset med sine 18.3 millioner ord er blant de mindre sammenlignet med Bank of English eller British National Corpus. Jo større et korpus er, desto mer sannsynlig er det at det dekker flere fenomener i språket. I tillegg til at vi ikke har tilgang på et mer omfattende korpus på norsk, er det også problematisk for NSP å prosessere ennå større mengder med tekst. Allerede nå måtte vi redusere korpuset til 15 millioner ord.

Tabellen viser at noen av testene gir forholdsvis like resultater. Spesielt PMI, Phi-testen og kjikvadrattesten synes å ha foretrukket de samme forekomstene på øverste plass. Fishertesten skiller seg mest ut, på grunn av at verditilordningen er 1 for en mengde på 71 kollokater i dette tilfellet. Denne testen beregner sannsynligheten for at det finnes en annen kontingenstabell med de samme marginale verdiene som det aktuelle n-grammet vi ser på. Sannsynligheten blir maksimal dersom det kan vises at en slik tabell finnes. Denne måten å rangere på gir et skjevt utslag når vi ser på så få linjer som i tabell 5.1. Hvilke av de bigrammene som listes med verdi 1 som kommer øverst i lista beror kun på tilfeldigheter. Til tross for at utviklerne av NSP anbefaler denne testen (Pedersen 1996), er det altså ikke åpenbart at den gir bedre resultater enn de asymptotiske i alle tilfeller.

5.1.2 Resultater fra fullformskorpus

For et søk i fullformer må brukeren spesifisere at hun vil ha treff for alle formene av ordet kaste, altså mengden {*kaste, kaster, kastet, kast*}. Trefflisten blir derfor lenger enn om vi bare hadde søkt på grunnformen *kaste*. Ettersom vi er ute etter distinkte ord som kan tilordnes *kaste*, er de forskjellige formene som repeterer samme bigram fjernet i denne tabellen. Vi

[3] Se også i tillegg B.1 for flere treff.

[4] Ifølge historien blir menn tvunget til å bære penisholdere, såkalte PH-er. Å kaste PH-en på bålet er "mannesaksmennenes" symbolske opprør mot de herskende kvinnene.

ser kun den første forekomsten av ordpar som kun avviker fra hverandre i bøyningsformen. Av totalt 383 forekomster med *kaste* som første ord, har vi her de første 35 tilordningene.

Plass	Dice	LeftFisher	Phi	PMI	X^2
1	lys	mig	smutt	smutt	smutt
2	bort	dem	loss	loss	loss
3	skygge	den	lys	blår	lys
4	stein	et	blår	stjålne	blår
5	loss	lys	sigarettstumpen	sigarettstumpen	stjålne
6	stjålne	glans	stjålne	vrak	sigarettstumpen
7	vrak	hodene	bort	snøballer	bort
8	lodd	lange	vrak	ballonger	vrak
9	glans	kortene	seg	regninger	et
10	smutt	jord	ballonger	småstein	lodd

Tabell 5.2: Kollokater til *kaste* i fullformsrangeringslister

De samme ordene som hadde en tydelig sterk assosiasjonsverdi i tabell 5.1 er høyt rangert også her. I tillegg kommer noen nye ord til som resultat av søk i fullformer. Dette kommer av assosiasjonsmålenes måter å utføre beregningene på. Hver form av ordet *kaste* blir betraktet som en selvstendig enhet, og ordene som forekommer i forbindelse med en form vil ikke ha noen påvirkning på ordene som forekommer sammen med en annen form av verbet. Effekten av dette kan være at noen n-grammer får en ufortjent lav assosiasjonsverdi fordi de forskjellige bøyningsformene sprer frekvenstellingene istedenfor å samle dem. Dette problemet unngår vi hvis vi søker på lemmaord. I de fleste tilfellene vil lemmaer derfor være å foretrekke.

5.2 Er objektiv evaluering mulig?

Det har vært foreslått mange måter å gi en tilnærmet objektiv evaluering av systemers kollokasjonsgjenkjenningsevne (f.eks. Evert 2004). En empirisk sammenlignende metode danner grunnlaget for en slik evaluering. En av de foreslåtte tilnærmingene er å benytte seg av rangerte resultatlister fra tekstkorpus, slike NSP returnerer, bestemme seg for et punkt i lista der man

antar at de fleste treffene er inneholdt, og beregne *presisjon* og *funnrate* for funnene før dette punktet. Denne tilnærmingen er en *semiautomatisk* prosess, ettersom den forutsetter en stor mengde med menneskelig forarbeid for å kunne gjennomføres. Til gjengjeld kan den gi en mer oversiktlig vurdering av assosiasjonsmålenes prestasjonsevne i forhold til å sammenligne enkelte utvalgte rangerte kandidater. Presisjon og funnrate (eng. "precision and recall") er de vanligste evalueringsmålene for søkemekanismer. Et annet mål som også tar høyde for rangering av resultatene er *ikke-interpolert gjennomsnittspresisjon* (Manning og Schütze 1999, s. 535).

5.2.1 Presisjon og funnrate

Ved å måle presisjon og funnrate vil vi ha svar på henholdsvis:

- Er den informasjonen vi har funnet relevant?
- Er all den relevante informasjonen funnet?

Vi kan representere informasjon i en toveis tabell:

	relevant	irrelevant
valgt	sp	fp
ikke valgt	fn	sn

Vi har et sett av elementer som vi ønsker skal bli identifisert. De elementene som systemet på korrekt måte har identifisert er plassert i ruten sp, som står for "sanne positiver" og sn, "sanne negativer". I rutene fp og fn finner vi henholdsvis "falske positiver", altså de som systemet feilaktig tok for å være i målgruppa, og "falske negativer", de elementene som ble valgt bort selv om de egentlig tilhører målgruppa.

Presisjon måler hvor stor andel av de utvalgte elementene som er relevante, og defineres som antall sanne postitive delt på det det totale antallet funn:

$$\text{Presisjon} = \frac{sp}{sp + fp}$$

Funnraten er den prosentuelle andelen av de relevante elementene som er funnet i forhold til alle relevante dokumenter i søkebasen:

$$\text{Funnrate} = \frac{sp}{sp + fn}$$

Problemet med å måle funnrate er at det er basert på at man vet på forhånd hvilke enheter som er relevante i henhold til noen veldefinerte kriterier. I oppgaven med å identifisere kollokasjoner er noe av problemet nettopp definisjonen av *hva* som er sanne positiver, altså virkelige kollokasjoner. Som vi så i kapittel 2, er det ikke åpenbart hvilke avgrensningskriterier vi kan bruke som grunnlag. Ettersom vi ikke har noen tilgjengelig standard/samling for norske kollokasjoner, har vi heller ingenting som kan tjene som et mål for *hvor mange* positive funn vi i utgangspunktet er på jakt etter. Det betyr også at vi bare kan måle presisjonen på systemet, ikke funnraten.

Målene presisjon og funnrate tar ikke høyde for at noen systemener returnerer dokumenter *rangert* etter relevans. NSP er et slikt system. Det gir faktisk alle n-grammer en rangeringsverdi, uansett om de er kollokasjonskandidater eller ikke. Systemet gir altså ikke et ja/nei-svar på om en ordkombinasjon er kollokasjon eller ikke, men plasserer den i en ordnet liste etter *graden* av sannsynlighet for at den er det. For å evaluere om systemet løser denne oppgaven på en tilfredsstillende måte, trenger vi et annet evalueringsmål som også bruker rangeringsinformasjon. I neste avsnitt presenteres et slikt mål.

5.2.2 Ikke-interpolert gjennomsnittspresisjon

Ikke-interpolert[5] gjennomsnittspresisjon ("uninterpolated average precision") er utviklet spesielt med tanke på å evaluere systemer som gir rangerte resultatlister. For rangerte mengder gir det liten mening å se på presisjonen alene. Løsningen i slike tilfeller er å etablere en terskel/avskjæring ("cutoff") for et visst antall funn vi ønsker å se på. Det er viktig å finne et avskjæringspunkt hvor man regner med at mesteparten av de relavante funnene er inkludert. Vi ønsker jo å slippe å vurdere mengdene av irrelevant materiale.

Ikke-interpolert gjennomsnittspresisjon (IGP) regnes ut ved å summere presisjonen for hvert enkelt punkt i listen hvor relevante funn er gjort og dele på mengden av relevante funn totalt. I formelle termer, hvor c er avskjæringspunktet og i det aktuelle n-grammet:

$$IGP = \frac{\sum_{i=1}^{c} \frac{sp_i}{sp_i + fp_i}}{sp_c}$$

[5]Interpolere: (mat.) beregne verdier som ligger mellom en serie verdier man kjenner.

5.2.3 Manuell annotering

For å kunne beregne presisjon og ikke-interpolert gjennomsnittspresisjon, må settet av utvunnede kollokasjonskandidater sammenlignes med en eller annen form for "gullstandard" som identifiserer kandidater som sanne positive (ekte kollokasjoner) og falske positive (ikke kollokasjoner). Kilder til slike referansedata kan være maskinelle ordbøker og leksika med søkbare ressurser for kollokasjonsinformasjon. For engelsk finnes mye tilgjengelig materiale for vurderingsgrunnlag. For eksempel Schone og Jurafsky (2001) bruker annet det såkalte WordNet med tilgang på 50 000 kollokasjoner (flerordsenheter), og en samling av internettordbøker til sine evalueringsoppgaver. Vi har beklageligvis ikke slike ressurser tilgjengelig for norsk, og må derfor etablere vår egen "gullstandard". Vi ønsker en presis definisjon av sanne positive, men kriteriene er ikke alltid åpenbare. Ved manuell gjennomgang av et utvalg bigrammer i absolutt rangering, har jeg forsøkt å holde meg til Sinclairs klassifikasjon og identifisere leksikalske objekter som kan plasseres i en av de fem kategoriene hans. Ideelt sett burde en gruppe på to eller flere kyndige lingvister eller leksikografer gått gjennom rangeringslistene ved å følge klare retningslinjer. Eksperters intuisjon spiller her en avgjørende rolle.

Resultatet av manuell annotering av rangeringene ble brukt til å sammenligne bigrammer fra tabeller med forskjellige størrelser på vinduene. Under dette arbeidet oppsto flere spørmål om tilhørighet som jeg mener Sinclairs klassifisering ikke dekker. Jeg har valgt å betrakte navn på lidelser (*multippel sklerose*), navn på matretter (*mango chutney*), militære uttrykk (*ballistisk missil*), tittel på institusjoner (*videregående skole*), datafaglige uttrykk (*logiske bombene*) og myntenheter (*portugisisk escudo*) som gyldige kollokasjoner utfra kriteriet om ikke-komposisjonalitet. Semantisk beslektede ord som *makrokosmos mikrokosmos* eller *anoreksi bulimi* er å regne for semantisk preferanse, så disse har også blitt merket av. Det finnes i tillegg en del utenlandske ord i teksten som også forekommer i fullformsordboka, og derfor blir representert i bigrammer. Ettersom de ikke har blitt filtrert ut, er det rimelig å gå ut fra at de er så mye brukt blant norsktalende at de kan betraktes som kollokasjonskandidater. Slike er f.eks. *bull shit* og *pro rata*. En annet potensielt hinder under manuell annotering er at mye materiale forutsetter detaljkunnskap i avgrensede fagområder. En stor mengde med juridiske og byråkratiske uttrykk krever et spesialisert ordforråd. Slike hindringer er nok en grunn til å involvere flere personer i det manuelle arbeidet.

5.3 Sammenligning av kontekststørrelser

Valg av størrelse på vinduet, spennvidden for et gitt n-gram har betydelig effekt på tellingen og rangeringen. For hver gang vi øker vindusstørrelsen i forhold til utgangspunktet, øker også antall n-grammer. For rangeringen betyr dette også at det er større sjanse for at n-grammer som er høyt rangert i fila med $k = 2$ kan havne mye lenger ned der $k = 3$.

For sammenligningens del har jeg beregnet vanlig presisjon og ikke-interpolert gjennomsnittspresisjon for rangeringen etter Dice-koeffisienten med bigrammer av lemmaord. Utregnigen krever manuell markering av sanne positiver. Avskjæringspunktet ble satt til 500.

Evalueringsmål	w2	w3	w4	w5
Presisjon	54 %	53 %	51 %	49 %
Ikke-interpolert gj.pres.	55 %	54 %	50 %	46 %

Tabell 5.3: Evalueringstall for distinkte vindusstørrelser ved Dice-koeffisient

Ifølge tabell 5.3 er både vanlig presisjon og ikke-interpolert gjennomsnittspresisjon bedre for mindre vindusspenn. Det er likevel viktig å huske på at tabellen kun viser et eksempel fra Dice-kvotientens rangeringsresultater, og at vurderingen av hvilke bigrammer som regnes som sanne positive er en subjektiv prosess. Vi kan også lese utfra tabellen at det tilsynelatende er tilfellet at presisjonen i en mengde på 500 holder seg relativt høy også ved større vindu. Det tyder på at antallet sanne positive fortsatt er høy i disse tilfellene (vindu 4 og 5), men at rangeringens kvalitet er fallende. Forklaringen på dette er delvis at mange gode kollokasjonskandidater blir "dyttet" lenger ned i lista ettersom nye bigrammer oppstår ved telling i en videre kontekst. Utvidelsen av konteksten medfører også en økning i antall bigrammer i datasettet. 500 er jo et svinnende lite antall i forhold til de nærmere 600 000 bigrammene med med $w = 2$ og 1 840 000 av de med $w = 5$! En riktigere måte for evaluering ville kanskje vært å øke avskjæringspunktet og vurderingsmengden ettersom det totale utvalget vokser. Arbeidet med å manuelt avmerke sanne positive er imidlertid så ressurskrevende at en slik oppgave blir overlatt til spesielt interesserte.

Det vi i alle fall *kan* se utfra evalueringsresultatene i tabell 5.3, er at over halvparten av bigrammene systemet returnerer er korrekt valgt ut som kollokasjonskandidater. Erfaring fra det manuelle arbeidet viser også at sanne positive ofte opptrer "puljevis" i rangeringen. Dette resulterer sannsynligvis av assosiasjonsmålets spesielle rangeringsegenskaper. Bigrammer som får lik

assosiasjonsverdi listes opp i vilkårlig rekkefølge innen denne verdimengden. Ofte deler ordpar som er gruppert sammen rundt samme verdi kollokasjonsegenskapene.

Fordelen med økning av kontekstvidde er at kollokasjoner som består av flere ord enn to vil ha muligheten til å få en høy verdi basert på de sentrale ordene. Et eksempel er *kort og godt*, et idiom som ikke er å finne i bigram med vindu 2, men som forekommer forholdsvis høyt oppe i rangeringen i bigramtelling med vindu 3.

5.4 Forslag til forbedringer

Selv om databasen `koll` nå ser ut til å inneholde mye nyttig informasjon som kan anvendes direkte i forhold til andre digitale leksikalske applikasjoner, er det mye som har potensiale til å bli bedre. Tids- og plassbegrensning i forbindelse med denne oppgaven gjør at disse punktene kun blir nevnt som forslag til videreutvikling av systemet.

1. Implementasjon av flere assosiasjonsmål for trigrammer. Det virker som om logaritmelikhetstesten ikke er det best egnede målet, men NSP inneholder ikke flere forslag til mål for trigrammer. Det hadde også vært interessant å implementere statistiske mål for n-grammer der $n > 3$. Ikke alle, men noen av assosiasjonsmålene kan generaliseres over også for flerdimensjonale tabeller. Det vil antakelig være en svært plasskrevende oppgave, slik at et mindre korpus er gunstig å bruke.

2. Mulig forbedring av resultatene for trigrammer kan oppnås ved å filtrere bort de hyppigste, men "uinteressante" bigrammene ("funksjonsordkombinasjoner"). Oppgaven er nok mer omfattende enn det ser ut til fordi vi vil ha en klar og systematisk definisjon av hvilke bigrammer dette gjelder. Det sier seg selv at ikke alle funksjonsord kan filtreres bort uten videre.

3. Implementasjon av en rangering der man kan krysse to forskjellige måleresultater hadde vært en gangbar måte for å oppnå høyere presisjon. Vi har gjort et forsøk på dette allerede ved bruk av skriptet `join.pl`. Foreløpig var det kun et eksperiment, men allerede den første versjonen så ut til å vise gode resultater. Igjen er det tiden som setter grenser for en mer systematisk utvikling av denne funksjonen.

Kapittel 6

Konklusjon

Som det ble nevnt innledningsvis, var målet med denne oppgaven å presentere en gangbar vei for automatisk identifisering og ekstrahering av kollokasjoner i det norske språket. Vi har sett at det *er* mulig å utvinne informasjon om assosiasjon mellom ord, og til og med gjenkjenne idiomer og faste uttrykk i språket utelukkende ved bruk av statistisk funderte metoder. Resultatene som NSP produserer er ikke dårlige tatt i betraktning at det ikke er mange flere arbeider som har vært viet denne oppgaven med norsk som fokusspråk.

Mange av vanskelighetene, spesielt med tanke på evaluering, har stammet fra mangelen på resultater etter forskningsarbeider innen samme felt. Evalueringsoppgaven har vist seg å være svært språkspesifikk, og erfaringer fra forskning på kollokasjoner i andre språk (tysk og engelsk) har ikke alltid vært mulig å adaptere direkte, til tross for at disse språkene er nært beslektet med norsk. Resultatene oppgaven kan rapportere om må betraktes som et forsøk på å etablere en form for "gullstandard", der andre interesserte kan ta opp igjen tråden og utbedre denne.

I maskinoversettelse innen LOGON-prosjektet kan identifiserte kollokasjoner anvendes direkte. Jeg ser for meg at kollokasjonsdatabasen kan integreres i kildespråkets leksikon. Under innlesning av norsk tekst kan denne databasen danne en aktiv komponent for en semantisk formalisering av enhetene bestående av flere ord. I "transfer"-delen produseres så formelle representasjoner som igjen danner grunnlag for generering av tilsvarende leksikalske enheter i målspråket. Løsninger for dette arbeidet blir ikke diskutert her, men overlatt til prosjektets faste medlemmer. Selv om det er sannsynlig at utbedringer av systemet vil være nødvendige i en konkret anvendelse, kan vi likevel konkludere med at grunnlaget for en integrasjon av flerordsenheter i det anvendte leksikonet er lagt. Det er fullt mulig å

videreutvikle og forfine systemet til å oppnå høyere presisjon.

Systemet kan også tjene som verktøy til hjelp ved visse leksikografiske oppgaver. For å lage et leksikon eller en spesialisert ordbok for idiomer og kollokasjoner kreves mye manuelt arbeid. Bruk av statistiske metoder kan forenkle jobben og minske ressursbruken for slike oppgaver. En enkel, men effektiv metode vil for eksempel være å samle, på systematisk måte, noen av de første n-grammene i rangeringen fra en eller flere av assosiasjonsmålene for hvert oppslagsord i en ordbok. Treffene kan gi et godt utgangspunkt for den manuelle konstruksjonen av en kollokasjonsordbok. For engelsk er Oxford Collocations Dictionary for English Learners er et utmerket eksempel på en slik samling. Leksika av denne typen er nyttig ikke minst i språklæringsprosesser.

Tillegg A

Utdrag fra korpus

I tabell A.1 ser vi korte eksempler på hvordan forskjellene mellom fullformskorpus og lemmatisert korpus arter seg. Det forekommer små uoverensstemmelser, som stor bokstav i begynnelsen av noen ord, tankestrek og lignende. I rad 5 er for eksempel ordet *skrevet* feilaktig blitt redusert til formen *skreve*. Slike feil genereres av taggeren og disambigueringsenheten ved Oslo-korpuset. De påvirker naturlig nok n-gramfrekvensene og de statistiske resultatene. Legg også merke til `_STOPP`-merkene som indikerer hvilke elementer vi skal filtrere bort under telling av n-grammer.

Fullformskorpus	**Lemmatisert korpus**
det har ikke alltid vært sånn	det ha ikke alltid være sånn
innholdet i tekstene har tatt ulike avstikkere	innhold i tekst ha ta ulik avstikker
tekstene på platene tar utgangspunkt i de små tingene men forhåpentligvis handler de om litt mer enn akkurat det	-tekst på plate ta utgangspunkt i de liten ting men forhåpentligvis handle de om litt mye enn akkurat det
de er konkrete og historiebaserte	De være konkret og historiebasere
jeg har skrevet min porsjon samfunnsrefsende harmdirrende tekster om krig og regnskog også	jeg ha skreve min porsjon samfunnsrefse harmdirrende tekst om krig og regnskog også
jeg har prøvd å være Bob_STOPP Dylan_STOPP og Ole_STOPP Paus_STOPP	jeg ha prøve å være Bob_STOPP Dylan_STOPP og Ole_STOPP Paus_STOPP
jeg har vært gjennom de fasene der jeg prøvde å kopiere ting jeg liker	jeg ha være gjennom de fase der jeg prøve å kopiere ting jeg like
det er vel sånn man lærer å skrive	det være vel sånn man lære å skrive
derfor er jeg utrolig glad jeg ikke fikk platekontrakt som attenåring	derfor være jeg utrolig glad jeg ikke få platekontrakt som attenåring
da jeg var atten var jeg uhyre skråsikker og idealistisk så det er like greit at tekstene fra den tiden ikke ble for offentlige	da jeg være atten være jeg uhyre skråsikker og idealistisk så det være like grei at tekst fra den tid ikke bli for offentlig

Tabell A.1: Eksempel på korpusformat i fullformer og lemmaer

Tillegg B

Utdrag fra rangeringer

I de neste avsnittene presenteres noen av de første treffene i rangeringene ordnet etter assosiasjonsmålenes resultater. I avsnitt B.1 er tabeller for rangeringen til n-grammer rundt ett utvalgt ord. Det neste avsnittet viser absolutte rangeringer i fullformer, uten noe spesielt søkeord. De siste tabellene må kun betraktes som eksempler på rangerte ordpar, der kollokasjonskandidater forekommer. Det kan være misvisende å prøve å trekke slutninger utfra et utvalg på 35 samforekomster når det totale antallet bigrammer er nærmere 700 000. Alle tabellene er basert på søk gjort over n-grammer bestående av ord som forekommer umiddelbart ved siden av hverandre, dvs. der vindusstørrelsen er 2 for bigrammer og 3 for trigrammer.

B.1 Direkte søk

Under følger de 35 første resultatene på søk i databasen på ordet *kaste*, sortert etter hvert av assosiasjonsmålene for fullformer og lemmaer, henholdsvis.

B.1.1 Fullformer

Dice-koeffisient

form1	form2	freq	dice
kaste	lys	41	0.0295
kaste	bort	49	0.0155
kaster	skygge	7	0.0149
kaste	stein	9	0.0133
kaster	loss	3	0.0113
kaster	stjålne	3	0.011
kaste	vrak	4	0.0108
kaste	lodd	4	0.0101
kaster	glans	3	0.0093
kaste	smutt	3	0.0085
kaster	ranselen	2	0.007
kaster	småstein	2	0.007
kastet	sig	18	0.0061
kaster	lampen	2	0.006
kaste	blår	2	0.0057
kastet	lange	10	0.0056
kaster	hodene	2	0.0056
kaste	regninger	2	0.0055
kaste	terninger	2	0.0053
kaste	band	2	0.0052
kaste	ball	2	0.0051
kastet	et	228	0.0051
kaste	mynt	2	0.005
kaste	klærne	3	0.0049
kaste	sten	2	0.0048
kaste	masken	2	0.0046
kaste	emnet	2	0.0045
kaste	dom	7	0.0044
kastet	seg	174	0.004
kastet	sigarettstumpen	3	0.004
kastet	spyd	3	0.0038
kastet	utfor	3	0.0037
kastet	ut	68	0.0037
kastet	kortene	3	0.0037

Venstresidig Fisher-test

form1	form2	freq	leftFisher
kastet	mig	3	1
kaster	dem	8	1
kaster	den	21	1
kaster	et	74	1
kastet	lys	5	1
kaster	glans	3	1
kaster	hodene	2	1
kastet	lange	10	1
kastet	kortene	3	1
kastet	jord	4	1
kaster	jeg	14	1
kaster	lampen	2	1
kaster	loss	3	1
kaster	lys	12	1
kaster	deg	5	1
kaster	bort	6	1
kastet	masken	3	1
kaste	regninger	2	1
kaste	seg	90	1
kaste	sig	8	1
kastet	meg	29	1
kaste	skygge	2	1
kaste	smutt	3	1
kaste	stein	9	1
kaste	sten	2	1
kaste	terninger	2	1
kaste	tunge	2	1
kaste	ut	10	1
kaste	vrak	4	1
kaste	ytterligere	2	1
kaster	meg	16	1
kaster	n	3	1
kastet	han	49	1
kastet	ham	17	1
kastet	ballen	2	1

Phi-test

form1	form2	freq	phi
kaste	smutt	3	0.0026
kaster	loss	3	0.0017
kaste	lys	41	0.0011
kaste	blår	2	0.0011
kastet	sigarettstumpen	3	0.0007
kaster	stjålne	3	0.0007
kaste	bort	49	0.0006
kaste	vrak	4	0.0005
kastet	et	228	0.0004
kastet	seg	174	0.0002
kastet	ballonger	2	0.0002
kaste	lodd	4	0.0002
kastet	snøballer	2	0.0002
kaste	regninger	2	0.0002
kaste	stein	9	0.0002
kaster	skygge	7	0.0002
kast	med	73	0.0002
kastet	harpunen	2	1e-04
kastet	overbord	2	1e-04
kastet	håndvesken	2	1e-04
kastet	bort	22	1e-04
kaste	seg	90	1e-04
kastet	foraktelig	2	1e-04
kaster	glans	3	1e-04
kastet	gjenskinn	2	1e-04
kaste	opp	50	1e-04
kastet	kortene	3	1e-04
kaste	terninger	2	1e-04
kastet	PH	2	1e-04
kaste	band	2	1e-04
kaste	ball	2	1e-04
kaste	mynt	2	1e-04
kastet	ut	68	1e-04
kastet	spyd	3	1e-04
kaster	småstein	2	1e-04

PMI

form1	form2	freq	pmi
kaste	smutt	3	13.2925
kaster	loss	3	12.7117
kaste	blår	2	12.7075
kaster	stjålne	3	11.3897
kastet	sigarettstumpen	3	11.3447
kaste	vrak	4	10.57
kastet	snøballer	2	10.1223
kastet	ballonger	2	10.0227
kaste	regninger	2	9.942
kaster	småstein	2	9.7762
kaster	ranselen	2	9.7482
kaste	lodd	4	9.4445
kaste	terninger	2	9.357
kaster	glans	3	9.1267
kastet	PH	2	9.0227
kaste	band	2	9.0071
kastet	håndvesken	2	8.6817
kastet	overbord	2	8.6442
kastet	gjenskinn	2	8.6077
kaste	ball	2	8.5865
kaster	skygge	7	8.5555
kaste	mynt	2	8.4445
kastet	harpunen	2	8.4061
kaste	lys	41	8.3619
kastet	foraktelig	2	8.3149
kaster	lampen	2	8.2687
kastet	spyd	3	8.1927
kastet	forte	2	8.1747
kaste	sten	2	7.9526
kaster	hodene	2	7.8487
kaste	stein	9	7.8374
kastet	kortene	3	7.7202
kaste	masken	2	7.6286
kastet	ballen	2	7.5721
kastet	utfor	3	7.5034

Kjikvadrattest

form1	form2	freq	x2
kaste	smutt	3	30095.1
kaster	loss	3	20118.7
kaste	lys	41	13409.5
kaste	blår	2	13374.2
kaster	stjålne	3	8043.91
kastet	sigarettstumpen	3	7797.02
kaste	bort	49	7059.02
kaste	vrak	4	6073.04
kastet	et	228	4169.7
kaste	lodd	4	2779.16
kaster	skygge	7	2619.96
kast	med	73	2551.02
kastet	seg	174	2450.82
kastet	snøballer	2	2225.44
kastet	ballonger	2	2076.81
kaste	stein	9	2040.74
kaste	regninger	2	1963.4
kaster	småstein	2	1749.8
kaster	ranselen	2	1716.07
kaster	glans	3	1671.08
kaste	seg	90	1423.73
kaste	opp	50	1367.91
kaste	terninger	2	1307.6
kaster	et	74	1248.83
kastet	PH	2	1036.41
kaste	band	2	1025.1
kastet	sten	4	916.829
kastet	ut	68	882.356
kastet	spyd	3	871.855
kastet	håndvesken	2	817.377
kastet	overbord	2	796.317
kastet	gjenskinn	2	776.309
kaste	ball	2	764.875
kaste	mynt	2	692.795
kastet	harpunen	2	674.531

Logaritmelikhet for trigrammer

id	form1	form2	form3	freq	ll
11313	å	kaste	seg	41	20424.2
113141	kaste	så	mye	7	13336.8
274844	han	kaste	seg	3	7730.26
20024	å	kaste	opp	27	6914.55
55473	å	kaste	en	12	4885.04
341642	kaste	den	første	3	4818.41
267435	å	kaste	ut	4	4783.85
302303	å	kaste	på	3	4760.78
398428	kaste	seg	til	3	4492.67
606383	kaste	seg	om	2	4304.5
110595	å	kaste	meg	7	4199.73
91394	kaste	seg	inn	8	4115.3
29276	kaste	seg	over	20	4020.34
40600	å	kaste	et	16	3892.08
746937	å	kaste	av	2	3768.38
343485	kaste	seg	ned	3	3668.03
611580	å	kaste	dem	2	3493.34
225153	å	kaste	henne	4	3237.56
349621	å	kaste	ham	3	3146.65
89819	og	kaste	seg	8	3066.26
20784	kaste	seg	ut	26	2910.23
92136	å	kaste	den	8	2812.82
427563	å	kaste	med	2	2781.43
331544	å	kaste	deg	3	2684.17
300788	å	kaste	sin	3	2479.33
168250	å	kaste	oss	5	2335.4
19585	å	kaste	bort	28	2329.93
31752	å	kaste	lys	19	2057.22
204311	å	kaste	i	4	2029.64
690153	å	kaste	sig	2	2004.2
44843	kaste	et	blikk	15	1970.42
255925	kaste	seg	på	4	1967.92
192059	å	kaste	det	5	1887.71
131311	kaste	seg	i	6	1884.86
164328	og	kaste	på	5	1871.76

B.1.2 Lemmaer

Dice-koeffisient

lemma1	lemma2	freq	dice
kaste	bort	82	0.0197
kaste	lys	62	0.0188
kaste	skygge	18	0.0092
kaste	stein	17	0.0087
kaste	sig	30	0.0085
kaste	seg	313	0.0071
kaste	opp	108	0.0069
kaste	ut	98	0.0051
kaste	loss	7	0.0051
kaste	snøball	7	0.0051
kaste	glans	7	0.0049
kaste	sten	7	0.0047
kaste	smutt	5	0.0037
kaste	terning	5	0.0036
kaste	vrak	5	0.0035
kaste	hode	15	0.0035
kaste	lodd	5	0.0035
kaste	maske	5	0.0032
kaste	lampe	5	0.0032
kaste	sneip	4	0.0029
kaste	tilbake	16	0.0029
kaste	harpun	4	0.0029
kaste	gjenskinn	4	0.0029
kaste	ball	4	0.0028
kaste	spyd	4	0.0028
kaste	søppel	4	0.0028
kaste	dom	8	0.0027
kaste	en	404	0.0027
kaste	blikk	8	0.0025
kaste	jakke	4	0.0025
kaste	stjålen	3	0.0022
kaste	PH	3	0.0022
kaste	ned	16	0.0022
kaste	blår	3	0.0022
kaste	overbord	3	0.0022

Venstresidig Fisher-test

lemma1	lemma2	freq	leftFisher
kaste	fram	9	1
kaste	hun	43	1
kaste	håndveske	2	1
kaste	lodd	5	1
kaste	sin	26	1
kaste	sigarettstump	2	1
kaste	inn	24	1
kaste	jakke	4	1
kaste	jeg	94	1
kaste	hode	15	1
kaste	harpun	4	1
kaste	hanske	3	1
kaste	fryktsom	2	1
kaste	småstein	2	1
kaste	smutt	5	1
kaste	gjenskinn	4	1
kaste	glans	7	1
kaste	skygge	18	1
kaste	guttunge	2	1
kaste	han	77	1
kaste	jord	5	1
kaste	kapitalisme	2	1
kaste	pille	2	1
kaste	maske	5	1
kaste	PH	3	1
kaste	overbord	3	1
kaste	over	19	1
kaste	murstein	2	1
kaste	opp	108	1
kaste	mynt	3	1
kaste	ransel	2	1
kaste	lys	62	1
kaste	lp	2	1
kaste	klær	4	1
kaste	sig	30	1

Phi-test

lemma1	lemma2	freq	phi
kaste	smutt	5	0.0018
kaste	loss	7	0.0018
kaste	blår	3	0.0007
kaste	snøball	7	0.0006
kaste	lp	2	0.0005
kaste	lys	62	0.0004
kaste	seg	313	0.0004
kaste	bort	82	0.0004
kaste	epleskrott	2	0.0002
kaste	sten	7	1e-04
kaste	stein	17	1e-04
kaste	en	404	1e-04
kaste	sneip	4	1e-04
kaste	skygge	18	1e-04
kaste	opp	108	1e-04
kaste	lodd	5	1e-04
kaste	sigarettstump	2	1e-04
kaste	overbord	3	1e-04
kaste	gjenskinn	4	1e-04
kaste	glans	7	1e-04
kaste	harpun	4	1e-04
kaste	stjålen	3	1e-04
kaste	sig	30	1e-04
kaste	ut	98	1e-04
kaste	vrak	5	1e-04
kaste	PH	3	1e-04
kaste	terning	5	1e-04
kaste	ned	16	0
kaste	over	19	0
kaste	mor	2	0
kaste	min	4	0
kaste	penge	3	0
kaste	mig	3	0
kaste	med	17	0
kaste	pille	2	0

PMI

lemma1	lemma2	freq	pmi
kaste	smutt	5	12.0789
kaste	loss	7	11.5644
kaste	lp	2	11.494
kaste	blår	3	11.342
kaste	snøball	7	10.0789
kaste	epleskrott	2	9.909
kaste	sigarettstump	2	8.757
kaste	PH	3	8.5764
kaste	overbord	3	8.3785
kaste	sneip	4	8.2716
kaste	fryktsom	2	8.1247
kaste	stjålen	3	8.1093
kaste	murstein	2	8.0345
kaste	terning	5	7.9579
kaste	gjenskinn	4	7.8891
kaste	harpun	4	7.831
kaste	vrak	5	7.6594
kaste	glans	7	7.5825
kaste	lodd	5	7.2716
kaste	foraktelig	2	7.246
kaste	småstein	2	7.172
kaste	søppel	4	7.1364
kaste	ball	4	7.1131
kaste	håndveske	2	7.1016
kaste	spyd	4	7.0021
kaste	mynt	3	6.9634
kaste	ransel	2	6.9496
kaste	sten	7	6.909
kaste	balle	3	6.6639
kaste	forte	2	6.6195
kaste	hanske	3	6.5346
kaste	ballong	2	6.494
kaste	kapitalisme	2	6.3924
kaste	band	2	6.3375
kaste	lys	62	6.1079

Kjikvadrattest

lemma1	lemma2	freq	x2
kaste	smutt	5	21626.6
kaste	loss	7	21189.9
kaste	blår	3	7783.18
kaste	snøball	7	7558.83
kaste	lp	2	5765.76
kaste	bort	82	5021.66
kaste	seg	313	4408.26
kaste	lys	62	4155.49
kaste	epleskrott	2	1919.26
kaste	en	404	1647.63
kaste	opp	108	1555.62
kaste	glans	7	1328.07
kaste	terning	5	1233.51
kaste	sneip	4	1228.39
kaste	skygge	18	1154.58
kaste	PH	3	1139.48
kaste	stein	17	1021.72
kaste	vrak	5	1001.09
kaste	overbord	3	992.623
kaste	ut	98	979.221
kaste	gjenskinn	4	940.475
kaste	harpun	4	903.036
kaste	sigarettstump	2	861.469
kaste	sig	30	832.109
kaste	sten	7	827.496
kaste	stjålen	3	822.649
kaste	lodd	5	762.775
kaste	søppel	4	554.936
kaste	fryktsom	2	554.372
kaste	ball	4	545.93
kaste	murstein	2	520.532
kaste	spyd	4	504.903
kaste	mynt	3	368.506
kaste	foraktelig	2	299.685
kaste	balle	3	298.296

Logaritmelikhet for trigrammer

id	lemma1	lemma2	lemma3	freq	ll
75778	kaste	inn	i	11	36443
13771	å	kaste	seg	41	20349.4
536450	kaste	den	ene	2	16827.7
32980	kaste	ut	av	21	15577.8
803294	kaste	en	gang	2	13476.6
763972	kaste	ut	fra	2	13357
619685	kaste	på	en	2	10976.9
35750	jeg	kaste	jeg	20	10878.3
147253	kaste	en	ny	6	9444.25
480911	kaste	i	den	2	8620.1
15717	han	kaste	seg	37	8582.98
41611	ha	kaste	seg	17	8112.31
106978	kaste	så	mye	8	8016.21
109069	så	kaste	han	8	7872.25
91588	kaste	ned	i	9	7297.63
860021	kaste	rest	av	2	7077.41
50283	hun	kaste	seg	15	7052.77
567450	kaste	jeg	ikke	2	6845.54
22468	å	kaste	en	28	6839.03
589720	kaste	han	ikke	2	6735.75
23784	å	kaste	opp	27	6196.71
295445	kaste	inn	på	4	5833.39
836121	kaste	det	av	2	5663.71
135978	kaste	på	hode	7	5632.37
267399	han	kaste	ikke	4	5398.75
133480	kaste	seg	til	7	5380.75
6442	kaste	seg	over	72	5248.24
91748	kaste	seg	om	9	5186.61
680453	kaste	ned	på	2	5089.43
505094	så	kaste	jeg	2	5044.27
23586	kaste	seg	inn	27	5000.98
95907	kaste	opp	i	9	4988.23
465815	kaste	ikke	engang	2	4912.56
486949	kaste	en	liten	2	4861.13
252280	kaste	den	første	4	4849.14

B.2 Absolutte rangeringer

Dice

form1	form2	dice
trom	trom	1
HOBBY	INTERIØR	1
meierismøret	sylten	1
visá	ávis	1
nam	nam	1
juksemaker	pipelort	1
tastafon	konvent	1
prosentlønnet	serveringspersonale	0.9524
inerte	anoder	0.88
skrutvinger	jigger	0.8571
siselerte	metallarbeider	0.8
nedgravet	frostfritt	0.8
sluttede	reiseselskaper	0.8
barnepsykiatriske	behandlingshjem	0.8
anabole	steroider	0.8
muggete	kornband	0.8
spesialtilpasset	båtkonstruksjoner	0.8
pro	rata	0.7907
fossile	brensler	0.7857
skålpund	gressfrø	0.7692
kvikksølvholdige	termometre	0.7692
oslo	infrastrukturinvesteringer	0.7692
ele	menter	0.75
aborterte	fostre	0.7383
avlangt	høydedrag	0.725
utadvendthet	kontaktevne	0.7143
kongelig	resolusjon	0.6929
multippel	sklerose	0.6897
statistisk	sentralbyrå	0.6894
provoserte	aborter	0.6849
generisk	substitusjon	0.678
ranke	giraffen	0.6667
nedbrytbar	herdeplast	0.6667
havbeiteprogrammet	push	0.6667
baconet	meierismøret	0.6667

LeftFisher

form1	form2	leftFisher
opp	i	1.0016
inn	i	1
han	hadde	1
å	få	1
å	være	1
slik	at	1
selv	om	1
forhold	til	1
når	det	1
i	dag	1
har	vært	1
i	forhold	1
ut	av	1
hun	hadde	1
å	bli	1
dette	er	1
seg	selv	1
å	ha	1
å	gjøre	1
kan	være	1
sammen	med	1
i	denne	1
å	se	1
det	gjelder	1
å	ta	1
ved	å	1
del	av	1
vi	har	1
hadde	vært	1
å	gå	1
sa	han	1
en	gang	1
men	jeg	1
i	tillegg	1
jeg	hadde	1

Phi-test

form1	form2	phi
trom	trom	1
HOBBY	INTERIØR	1
meierismøret	sylten	1
visá	ávis	1
nam	nam	1
juksemaker	pipelort	1
tastafon	konvent	1
prosentlønnet	serveringspersonale	0.9091
inerte	anoder	0.7756
skrutvinger	jigger	0.75
siselerte	metallarbeider	0.6667
nedgravet	frostfritt	0.6667
sluttede	reiseselskaper	0.6667
barnepsykiatriske	behandlingshjem	0.6667
anabole	steroider	0.6667
muggete	kornband	0.6667
spesialtilpasset	båtkonstruksjoner	0.6667
pro	rata	0.6538
skålpund	gressfrø	0.625
oslo	infrastrukturinvesteringer	0.625
fossile	brensler	0.6185
ele	menter	0.6
kvikksølvholdige	termometre	0.5952
aborterte	fostre	0.568
avlangt	høydedrag	0.5286
statistisk	sentralbyrå	0.5224
utadvendthet	kontaktevne	0.5208
havbeiteprogrammet	push	0.5
baconet	meierismøret	0.5
rikk	rikk	0.5
dampdrevne	elektrisitetsverket	0.5
stikkprøvekontroll	fylkesoversikt	0.5
reversible	varmeeffekten	0.5
kongelig	resolusjon	0.495
multippel	sklerose	0.4808

PMI

form1	form2	pmi
tastafon	konvent	22.4765
meierismøret	sylten	22.4765
visá	ávis	22.4765
barnepsykiatriske	behandlingshjem	21.8916
anabole	steroider	21.8916
muggete	kornband	21.8916
spesialtilpasset	båtkonstruksjoner	21.8916
juksemaker	pipelort	21.8916
sluttede	reiseselskaper	21.8916
siselerte	metallarbeider	21.8916
nedgravet	frostfritt	21.8916
HOBBY	INTERIØR	21.8916
nam	nam	21.8916
dampdrevne	elektrisitetsverket	21.4765
stikkprøvekontroll	fylkesoversikt	21.4765
baconet	meierismøret	21.4765
trom	trom	21.4765
rikk	rikk	21.4765
skrutvinger	jigger	21.4765
arbeidsbesparende	husholdningsmaskiner	21.3066
nedbrytbar	herdeplast	21.3066
kapslete	likestrøm	21.1546
svovelsyre	flussyre	21.1546
ele	menter	21.1546
revmatologisk	immunologi	21.1546
asbest	asbestholdige	21.1546
reversible	varmeeffekten	20.8916
anne	kjersti	20.8916
havbeiteprogrammet	push	20.8916
atlantiske	lavtrykk	20.8916
breivik	botn	20.8916
uberettiga	inntrenging	20.6692
fasetterte	flakonger	20.6692
skvalpende	retorten	20.6692
osende	tranlampe	20.6692

Kjikvadrattest

form1	form2	x2
trom	trom	1.16718e+07
HOBBY	INTERIØR	1.16718e+07
meierismøret	sylten	1.16718e+07
visá	ávis	1.16718e+07
nam	nam	1.16718e+07
juksemaker	pipelort	1.16718e+07
tastafon	konvent	1.16718e+07
prosentlønnet	serveringspersonale	1.06107e+07
inerte	anoder	9.05314e+06
skrutvinger	jigger	8.75386e+06
siselerte	metallarbeider	7.78121e+06
nedgravet	frostfritt	7.78121e+06
sluttede	reiseselskaper	7.78121e+06
barnepsykiatriske	behandlingshjem	7.78121e+06
anabole	steroider	7.78121e+06
muggete	kornband	7.78121e+06
spesialtilpasset	båtkonstruksjoner	7.78121e+06
pro	rata	7.63156e+06
skålpund	gressfrø	7.29488e+06
oslo	infrastrukturinvesteringer	7.29488e+06
fossile	brensler	7.2188e+06
ele	menter	7.00309e+06
kvikksølvholdige	termometre	6.94751e+06
aborterte	fostre	6.63006e+06
avlangt	høydedrag	6.16969e+06
statistisk	sentralbyrå	6.09755e+06
utadvendthet	kontaktevne	6.07907e+06
baconet	meierismøret	5.83591e+06
rikk	rikk	5.83591e+06
dampdrevne	elektrisitetsverket	5.83591e+06
stikkprøvekontroll	fylkesoversikt	5.83591e+06
havbeiteprogrammet	push	5.83591e+06
reversible	varmeeffekten	5.83591e+06
kongelig	resolusjon	5.77729e+06
multippel	sklerose	5.61144e+06

Logaritmelikhet for trigrammer

form1	form2	form3	ll
til	å	få	183110
forhold	til	å	170594
til	å	gjøre	169493
til	å	være	167874
til	å	bli	166817
til	å	se	166055
til	å	ta	165966
til	å	gå	163299
til	å	gi	160658
til	å	ha	160642
til	å	finne	160035
til	å	si	159223
til	å	i	157814
til	å	komme	156749
til	å	holde	156409
knyttet	til	å	155317
til	å	tenke	154862
til	å	snakke	154218
til	å	sette	153142
til	å	bruke	152842
til	å	sikre	152521
hensyn	til	å	152254
til	å	legge	152227
til	å	å	151952
til	å	forstå	150464
til	å	skape	150310
til	å	unngå	149780
til	å	redusere	149752
til	å	vurdere	149672
til	å	oppnå	149603
til	å	ikke	149302
til	å	skaffe	149035
til	å	utvikle	149035
til	å	vise	149019
til	å	trekke	148984

Bibliografi

Banerjee, S. og Pedersen, T. (2003), The Design, Implementation and Use of the N-gram Statistic Package, *in* 'Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics', Mexico City.

Church, K. W. og Hanks, P. (1990), 'Word Assosiation Norms, Mutual Information, and Lexicography', *Computational Linguistics* **16**(1), 22–29.

Evert, S. (2004), The Statistics of Word Cooccurrenses - Word Pairs and Collocations, PhD thesis, Universität Stuttgart.

Firth, J. R. (1957), A synopsis of linguistic theory 1930-55, *in* F. R. Palmer, ed., 'Selected papers of J. R. Firth 1952-59', Longmans' Linguistic Library, pp. 168–205. Special volume of the Philological Society.

Kiss, T. og Strunk, J. (2002), Viewing sentence boundary detection as collocation identification, *in* S. Busemann, ed., 'Tagungsband der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)', Saarbrücken, Germany, pp. 75–82.

Løvås, G. G. (2004), *Statistikk for universiteter og høgskoler*, 2. edn, Universitetsforlaget.

Manning, C. D. og Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

Pedersen, T. (1996), Fishing for Exactness, *in* 'Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)', Austin, TX.

Rapp, R. (2002), The computation of word associations: Comparing syntagmatic and paradigmatic approaches, *in* 'Proceedings of COLING 2002', Taipeh, Taiwan.

Sapir, E. (1921), *Language: An introduction to the study of speech*, Harcourt Brace, New York.

Schone, P. og Jurafsky, D. (2001), Is knowledgefree induction of multiword unit dictionary headwords a solved problem?, *in* 'Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing', Pittsburgh, PA, pp. 100–108.

Sinclair, J. (1991), *Corpus, Concordance, Collocation*, Oxford University Press.

Sinclair, J. (1999), 'The Computer, the Corpus and the Theory of Language', *LMS Lingua* (1), 21–32. Riksföreningen för Lärarna i Moderna Språk.

Smadja, F., McKeown, K. R. og Hatzivassiloglou, V. (1996), 'Translating collocations for bilingual lexicons: A statistical approach', *Computational Linguistics* **22**(1), 1–38.

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA.